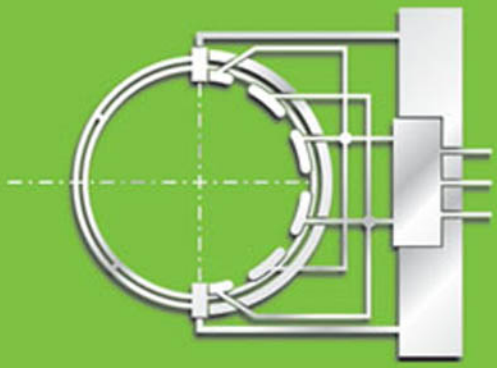
THIRD EDITION

HANDBOOK OF MODERN SENSORS

Physics, Designs, and Applications



JACOB FRADEN

VISIT…

HANDBOOK OF
MODERN SENSORS

PHYSICS, DESIGNS, and APPLICATIONS

Third Edition

Springer
New York
Berlin
Heidelberg
Hong Kong
London
Milan
Paris
Tokyo

HANDBOOK OF
MODERN SENSORS

PHYSICS, DESIGNS, and APPLICATIONS

Third Edition

JACOB FRADEN

Advanced Monitors Corporation
San Diego, California

With 403 lillustrations

AIP PRESS

Springer

Jacob Fraden
Advanced Monitors Corporation
6255 Ferris Square, Suite M
San Diego, CA 92121
USA
jfraden@admon.com

Library of Congress Cataloging-in-Publication Data
Fraden, Jacob
Handbook of modern sensors : physics, designs, and applications / Jacob Fraden.–3rd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-387-00750-4 (alk. paper)
1. Detectors–Handbooks, manuals, etc. 2. Interface circuits–Handbooks, manuals, etc.
I. Title.

TA165.F723 2003
681′.2—dc21 2003044597

ISBN 0-387-00750-4 Printed on acid-free paper.

AIP Press is an imprint of Springer-Verlag, Inc.

© 2004, 1996 Springer-Verlag New York, Inc.
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 10919477

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

To the memory of my father

This page intentionally left blank

Preface

Seven years have passed since the publication of the previous edition of this book. During that time, sensor technologies have made a remarkable leap forward. The sensitivity of the sensors became higher, the dimensions became smaller, the selectivity became better, and the prices became lower. What have not changed are the fundamental principles of the sensor design. They are still governed by the laws of Nature. Arguably one of the greatest geniuses who ever lived, Leonardo Da Vinci, had his own peculiar way of praying. He was saying, "*Oh Lord, thanks for Thou do not violate your own laws*." It is comforting indeed that the laws of Nature do not change as time goes by; it is just our appreciation of them that is being refined. Thus, this new edition examines the same good old laws of Nature that are employed in the designs of various sensors. This has not changed much since the previous edition. Yet, the sections that describe the practical designs are revised substantially. Recent ideas and developments have been added, and less important and nonessential designs were dropped. Probably the most dramatic recent progress in the sensor technologies relates to wide use of MEMS and MEOMS (*micro-electro-mechanical systems* and *micro-electro-opto-mechanical systems)*. These are examined in this new edition with greater detail.

This book is about devices commonly called sensors. The invention of a microprocessor has brought highly sophisticated instruments into our everyday lives. Numerous computerized appliances, of which microprocessors are integral parts, wash clothes and prepare coffee, play music, guard homes, and control room temperature. Microprocessors are digital devices that manipulate binary codes generally represented by electric signals. Yet, we live in an analog world where these devices function among objects that are mostly not digital. Moreover, this world is generally not electrical (apart from the atomic level). Digital systems, however complex and intelligent they might be, must receive information from the outside world. Sensors are interface devices between various physical values and electronic circuits who "understand" only a language of moving electrical charges. In other words, sensors are the eyes, ears, and noses of silicon chips. Sensors have become part of everyone's life. In the United States alone, they comprise a $12 billion industry.

In the course of my engineering work, I often felt a strong need for a book that would combine practical information on diversified subjects related to the most important physical principles, design, and use of various sensors. Surely, I could find almost all I had to know in texts on physics, electronics, technical magazines, and manufacturers' catalogs. However, the information is scattered over many publications, and almost every question I was pondering required substantial research work and numerous trips to the library. Little by little, I have been gathering practical information on everything that in any way was related to various sensors and their applications to scientific and engineering measurements. Soon, I realized that the information I collected might be quite useful to more than one person. This idea prompted me to write this book.

In setting my criteria for selecting various sensors for this edition, I attempted to keep the scope of this book as broad as possible, opting for brief descriptions of many different designs (without being trivial, I hope) rather than fewer treated in greater depth. This volume attempts (immodestly perhaps) to cover a very broad range of sensors and detectors. Many of them are well known, but describing them is still useful for students and those who look for a convenient reference. It is the author's intention to present a comprehensive and up-to-date account of the theory (physical principles), design, and practical implementations of various (especially the newest) sensors for scientific, industrial, and consumer applications. The topics included in the book reflect the author's own preferences and interpretations. Some may find a description of a particular sensor either too detailed or too broad or, contrary, too brief. In most cases, the author tried to make an attempt to strike a balance between a detailed description and a simplicity of coverage.

This volume covers many modern sensors and detectors. It is clear that one book cannot embrace the whole variety of sensors and their applications, even if it is called something like *The Encyclopedia of Sensors*. This is a different book, and the author's task was much less ambitious. Here, an attempt has been made to generate a reference text that could be used by students, researchers interested in modern instrumentation (applied physicists and engineers), sensor designers, application engineers, and technicians whose job is to understand, select, and/or design sensors for practical systems.

The previous editions of this book have been used quite extensively as desktop references and textbooks for the related college courses. Comments and suggestions from the sensor designers, professors, and students prompted me to implement several changes and correct errors.

Jacob Fraden
San Diego, California
November 2003

Contents

Preface VII

1 Data Acquisition 1
1.1 Sensors, Signals, and Systems 1
1.2 Sensor Classification 7
1.3 Units of Measurements 9
References 11

2 Sensor Characteristics 13
2.1 Transfer Function 13
2.2 Span (Full-Scale Input) 15
2.3 Full-Scale Output 16
2.4 Accuracy 17
2.5 Calibration 18
2.6 Calibration Error 19
2.7 Hysteresis 20
2.8 Nonlinearity 20
2.9 Saturation 22
2.10 Repeatability 23
2.11 Dead Band 23
2.12 Resolution 23
2.13 Special Properties 24
2.14 Output Impedance 24
2.15 Excitation 25
2.16 Dynamic Characteristics 25
2.17 Environmental Factors 29
2.18 Reliability 31
2.19 Application Characteristics 33
2.20 Uncertainty 33
References 35

3 Physical Principles of Sensing 37
3.1 Electric Charges, Fields, and Potentials 38
3.2 Capacitance 44
3.2.1 Capacitor 45
3.2.2 Dielectric Constant 46
3.3 Magnetism 50
3.3.1 Faraday's Law 52
3.3.2 Solenoid 54
3.3.3 Toroid 55
3.3.4 Permanent Magnets 55
3.4 Induction 56
3.5 Resistance 59
3.5.1 Specific Resistivity 60
3.5.2 Temperature Sensitivity 62
3.5.3 Strain Sensitivity 64
3.5.4 Moisture Sensitivity 65
3.6 Piezoelectric Effect 66
3.6.1 Piezoelectric Films 72
3.7 Pyroelectric Effect 76
3.8 Hall Effect 82
3.9 Seebeck and Peltier Effects 86
3.10 Sound Waves 92
3.11 Temperature and Thermal Properties of Materials 94
3.11.1 Temperature Scales 95
3.11.2 Thermal Expansion 96
3.11.3 Heat Capacity 98
3.12 Heat Transfer 99
3.12.1 Thermal Conduction 99
3.12.2 Thermal Convection 102
3.12.3 Thermal Radiation 103
3.12.3.1 Emissivity 106
3.12.3.2 Cavity Effect 109
3.13 Light 111
3.14 Dynamic Models of Sensor Elements 113
3.14.1 Mechanical Elements 115
3.14.2 Thermal Elements 117
3.14.3 Electrical Elements 118
3.14.4 Analogies 119
References 119

4 Optical Components of Sensors 123
4.1 Radiometry 125
4.2 Photometry 129
4.3 Windows 132
4.4 Mirrors 134

4.5 Lenses . . . 136
4.6 Fresnel Lenses . . . 137
4.7 Fiber Optics and Waveguides . . . 140
4.8 Concentrators . . . 144
4.9 Coatings for Thermal Absorption . . . 145
4.10 Electro-optic and Acousto-optic Modulators . . . 146
4.11 Interferometric Fiber-optic Modulation . . . 148
References . . . 149

5 Interface Electronic Circuits . . . 151
5.1 Input Characteristics of Interface Circuits . . . 151
5.2 Amplifiers . . . 156
5.2.1 Operational Amplifiers . . . 156
5.2.2 Voltage Follower . . . 158
5.2.3 Instrumentation Amplifier . . . 159
5.2.4 Charge Amplifiers . . . 161
5.3 Excitation Circuits . . . 164
5.3.1 Current Generators . . . 165
5.3.2 Voltage References . . . 169
5.3.3 Oscillators . . . 171
5.3.4 Drivers . . . 174
5.4 Analog-to-Digital Converters . . . 175
5.4.1 Basic Concepts . . . 175
5.4.2 V/F Converters . . . 176
5.4.3 Dual-Slope Converter . . . 181
5.4.4 Successive-Approximation Converter . . . 183
5.4.5 Resolution Extension . . . 185
5.5 Direct Digitization and Processing . . . 186
5.6 Ratiometric Circuits . . . 190
5.7 Bridge Circuits . . . 192
5.7.1 Disbalanced Bridge . . . 193
5.7.2 Null-Balanced Bridge . . . 194
5.7.3 Temperature Compensation of Resistive Bridge . . . 195
5.7.4 Bridge Amplifiers . . . 200
5.8 Data Transmission . . . 201
5.8.1 Two-Wire Transmission . . . 202
5.8.2 Four-Wire Sensing . . . 203
5.8.3 Six-Wire Sensing . . . 204
5.9 Noise in Sensors and Circuits . . . 204
5.9.1 Inherent Noise . . . 205
5.9.2 Transmitted Noise . . . 207
5.9.3 Electric Shielding . . . 212
5.9.4 Bypass Capacitors . . . 214
5.9.5 Magnetic Shielding . . . 215
5.9.6 Mechanical Noise . . . 217

5.9.7 Ground Planes . . . 218
5.9.8 Ground Loops and Ground Isolation . . . 219
5.9.9 Seebeck Noise . . . 221
5.10 Batteries for Low Power Sensors . . . 222
5.10.1 Primary Cells . . . 223
5.10.2 Secondary Cells . . . 224
References . . . 225

6 Occupancy and Motion Detectors . . . 227
6.1 Ultrasonic Sensors . . . 228
6.2 Microwave Motion Detectors . . . 228
6.3 Capacitive Occupancy Detectors . . . 233
6.4 Triboelectric Detectors . . . 237
6.5 Optoelectronic Motion Detectors . . . 238
6.5.1 Sensor Structures . . . 240
6.5.1.1 Multiple Sensors . . . 241
6.5.1.2 Complex Sensor Shape . . . 241
6.5.1.3 Image Distortion . . . 241
6.5.1.4 Facet Focusing Element . . . 242
6.5.2 Visible and Near-Infrared Light Motion Detectors . . . 243
6.5.3 Far-Infrared Motion Detectors . . . 244
6.5.3.1 PIR Motion Detectors . . . 245
6.5.3.2 PIR Sensor Efficiency Analysis . . . 247
References . . . 251

7 Position, Displacement, and Level . . . 253
7.1 Potentiometric Sensors . . . 254
7.2 Gravitational Sensors . . . 256
7.3 Capacitive Sensors . . . 258
7.4 Inductive and Magnetic Sensors . . . 262
7.4.1 LVDT and RVDT . . . 262
7.4.2 Eddy Current Sensors . . . 264
7.4.3 Transverse Inductive Sensor . . . 266
7.4.4 Hall Effect Sensors . . . 267
7.4.5 Magnetoresistive Sensors . . . 271
7.4.6 Magnetostrictive Detector . . . 274
7.5 Optical Sensors . . . 275
7.5.1 Optical Bridge . . . 275
7.5.2 Proximity Detector with Polarized Light . . . 276
7.5.3 Fiber-Optic Sensors . . . 278
7.5.4 Fabry–Perot Sensors . . . 278
7.5.5 Grating Sensors . . . 281
7.5.6 Linear Optical Sensors (PSD) . . . 283
7.6 Ultrasonic Sensors . . . 286
7.7 Radar Sensors . . . 289

7.7.1 Micropower Impulse Radar . . . 289
7.7.2 Ground-Penetrating Radar . . . 291
7.8 Thickness and Level Sensors . . . 293
7.8.1 Ablation Sensors . . . 293
7.8.2 Thin-Film Sensors . . . 296
7.8.3 Liquid-Level Sensors . . . 296
References . . . 298

8 Velocity and Acceleration . . . 301
8.1 Accelerometer Characteristics . . . 303
8.2 Capacitive Accelerometers . . . 305
8.3 Piezoresistive Accelerometers . . . 307
8.4 Piezoelectric Accelerometers . . . 309
8.5 Thermal Accelerometers . . . 309
8.5.1 Heated-Plate Accelerometer . . . 309
8.5.2 Heated-Gas Accelerometer . . . 310
8.6 Gyroscopes . . . 313
8.6.1 Rotor Gyroscope . . . 313
8.6.2 Monolithic Silicon Gyroscopes . . . 314
8.6.3 Optical Gyroscopes . . . 317
8.7 Piezoelectric Cables . . . 319
References . . . 321

9 Force, Strain, and Tactile Sensors . . . 323
9.1 Strain Gauges . . . 325
9.2 Tactile Sensors . . . 327
9.3 Piezoelectric Force Sensors . . . 334
References . . . 336

10 Pressure Sensors . . . 339
10.1 Concepts of Pressure . . . 339
10.2 Units of Pressure . . . 340
10.3 Mercury Pressure Sensor . . . 341
10.4 Bellows, Membranes, and Thin Plates . . . 342
10.5 Piezoresistive Sensors . . . 344
10.6 Capacitive Sensors . . . 349
10.7 VRP Sensors . . . 350
10.8 Optoelectronic Sensors . . . 352
10.9 Vacuum Sensors . . . 354
10.9.1 Pirani Gauge . . . 354
10.9.2 Ionization Gauges . . . 356
10.9.3 Gas Drag Gauge . . . 356
References . . . 357

11 Flow Sensors 359
11.1 Basics of Flow Dynamics 359
11.2 Pressure Gradient Technique 361
11.3 Thermal Transport Sensors 363
11.4 Ultrasonic Sensors 367
11.5 Electromagnetic Sensors 370
11.6 Microflow Sensors 372
11.7 Breeze Sensor 374
11.8 Coriolis Mass Flow Sensors 376
11.9 Drag Force Flow Sensors 377
References 378

12 Acoustic Sensors 381
12.1 Resistive Microphones 382
12.2 Condenser Microphones 382
12.3 Fiber-Optic Microphone 383
12.4 Piezoelectric Microphones 385
12.5 Electret Microphones 386
12.6 Solid-State Acoustic Detectors 388
References 391

13 Humidity and Moisture Sensors 393
13.1 Concept of Humidity 393
13.2 Capacitive Sensors 396
13.3 Electrical Conductivity Sensors 399
13.4 Thermal Conductivity Sensor 401
13.5 Optical Hygrometer 402
13.6 Oscillating Hygrometer 403
References 404

14 Light Detectors 407
14.1 Introduction 407
14.2 Photodiodes 411
14.3 Phototransistor 418
14.4 Photoresistors 420
14.5 Cooled Detectors 423
14.6 Thermal Detectors 425
14.6.1 Golay Cells 426
14.6.2 Thermopile Sensors 427
14.6.3 Pyroelectric Sensors 430
14.6.4 Bolometers 434
14.6.5 Active Far-Infrared Sensors 437
14.7 Gas Flame Detectors 439
References 441

15 Radiation Detectors 443
15.1 Scintillating Detectors 444
15.2 Ionization Detectors 447
15.2.1 Ionization Chambers 447
15.2.2 Proportional Chambers 449
15.2.3 Geiger–Müller Counters 450
15.2.4 Semiconductor Detectors 451
References 455

16 Temperature Sensors 457
16.1 Thermoresistive Sensors 461
16.1.1 Resistance Temperature Detectors 461
16.1.2 Silicon Resistive Sensors 464
16.1.3 Thermistors 465
16.1.3.1 NTC Thermistors 465
16.1.3.2 Self-Heating Effect in NTC Thermistors 474
16.1.3.3 PTC Thermistors 477
16.2 Thermoelectric Contact Sensors 481
16.2.1 Thermoelectric Law 482
16.2.2 Thermocouple Circuits 484
16.2.3 Thermocouple Assemblies 486
16.3 Semiconductor P-N Junction Sensors 488
16.4 Optical Temperature Sensors 491
16.4.1 Fluoroptic Sensors 492
16.4.2 Interferometric Sensors 494
16.4.3 Thermochromic Solution Sensor 494
16.5 Acoustic Temperature Sensor 495
16.6 Piezoelectric Temperature Sensors 496
References 497

17 Chemical Sensors 499
17.1 Chemical Sensor Characteristics 500
17.2 Specific Difficulties 500
17.3 Classification of Chemical-Sensing Mechanisms 501
17.4 Direct Sensors 503
17.4.1 Metal-Oxide Chemical Sensors 503
17.4.2 ChemFET 504
17.4.3 Electrochemical Sensors 505
17.4.4 Potentiometric Sensors 506
17.4.5 Conductometric Sensors 507
17.4.6 Amperometric Sensors 508
17.4.7 Enhanced Catalytic Gas Sensors 510
17.4.8 Elastomer Chemiresistors 512
17.5 Complex Sensors 512
17.5.1 Thermal Sensors 513

17.5.2 Pellister Catalytic Sensors 514
17.5.3 Optical Chemical Sensors 514
17.5.4 Mass Detector 516
17.5.5 Biochemical Sensors 519
17.5.6 Enzyme Sensors 520
17.6 Chemical Sensors Versus Instruments 520
17.6.1 Chemometrics 523
17.6.2 Multisensor Arrays 524
17.6.3 Electronic Noses (Olfactory Sensors) 524
17.6.4 Neural Network Signal (Signature) Processing for Electronic Noses 527
17.6.5 "Smart" Chemical Sensors 530
References 530

18 Sensor Materials and Technologies 533
18.1 Materials 533
18.1.1 Silicon as a Sensing Material 533
18.1.2 Plastics 536
18.1.3 Metals 540
18.1.4 Ceramics 542
18.1.5 Glasses 543
18.2 Surface Processing 543
18.2.1 Deposition of Thin and Thick Films 543
18.2.2 Spin-Casting 544
18.2.3 Vacuum Deposition 544
18.2.4 Sputtering 545
18.2.5 Chemical Vapor Deposition 546
18.3 Nano-Technology 547
18.3.1 Photolithography 548
18.3.2 Silicon Micromachining 549
18.3.2.1 Basic Techniques 549
18.3.2.2 Wafer bonding 554
References 555

Appendix 557
Table A.1 Chemical Symbols for the Elements 557
Table A.2 SI Multiples 558
Table A.3 Derivative SI Units 558
Table A.4 SI Conversion Multiples 559
Table A.5 Dielectric Constants of Some Materials at Room Temperature 564
Table A.6 Properties of Magnetic Materials 564
Table A.7 Resistivities and Temperature Coefficients of Resistivity of Some Materials at Room Temperature 565
Table A.8 Properties of Piezoelectric Materials at 20°C 565

Table A.9 Physical Properties of Pyroelectric Materials 566
Table A.10 Characteristics of Thermocouple Types 566
Table A.11 Thermoelectric Coefficients and Volume Resistivities of Selected Elements 567
Table A.11a Thermocouples for Very Low and Very High Temperatures 567
Table A.12 Densities of Some Materials 568
Table A.13 Mechanical Properties of Some Solid Materials 568
Table A.14 Mechanical Properties of Some Crystalline Materials 569
Table A.15 Speed of Sound Waves 569
Table A.16 Coefficient of Linear Thermal Expansion of Some Materials 569
Table A.17 Specific Heat and Thermal Conductivity of Some Materials 570
Table A.18 Typical Emissivities of Different Materials 571
Table A.19 Refractive Indices of Some Materials 572
Table A.20 Characteristics of C–Zn and Alkaline Cells 573
Table A.21 Lithium–Manganese Dioxide Primary Cells 573
Table A.22 Typical Characteristics of "AA"-Size Secondary Cells 573
Table A.23 Miniature Secondary Cells and Batteries 574
Table A.24 Electronic Ceramics 576
Table A.25 Properties of Glasses 577

Index 579

This page intentionally left blank

1

Data Acquisition

"It's as large as life, and twice as natural"

—Lewis Carroll, "Through the Looking Glass"

1.1 Sensors, Signals, and Systems

A sensor is often defined as a *device that receives and responds to a signal or stimulus*. This definition is broad. In fact, it is so broad that it covers almost everything from a human eye to a trigger in a pistol. Consider the level-control system shown in Fig. 1.1 [1]. The operator adjusts the level of fluid in the tank by manipulating its valve. Variations in the inlet flow rate, temperature changes (these would alter the fluid's viscosity and, consequently, the flow rate through the valve), and similar disturbances must be compensated for by the operator. Without control, the tank is likely to flood, or run dry. To act appropriately, the operator must obtain information about the level of fluid in the tank on a timely basis. In this example, the information is perceived by the sensor, which consists of two main parts: the sight tube on the tank and the operator's eye, which generates an electric response in the optic nerve. The sight tube by itself is not a sensor, and in this particular control system, the eye is not a sensor either. Only the combination of these two components makes a narrow-purpose sensor (detector), which is *selectively* sensitive to the fluid level. If a sight tube is designed properly, it will very quickly reflect variations in the level, and it is said that the sensor has a fast speed response. If the internal diameter of the tube is too small for a given fluid viscosity, the level in the tube may lag behind the level in the tank. Then, we have to consider a phase characteristic of such a sensor. In some cases, the lag may be quite acceptable, whereas in other cases, a better sight tube design would be required. Hence, the sensor's performance must be assessed only as a part of a data acquisition system.

This world is divided into natural and man-made objects. The natural sensors, like those found in living organisms, usually respond with signals, having an electrochemical character; that is, their physical nature is based on ion transport, like in the nerve fibers (such as an optic nerve in the fluid tank operator). In man-made devices,

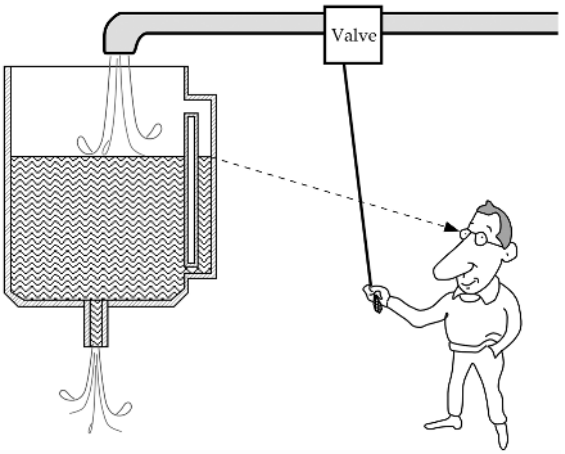


Fig. 1.1. Level-control system. A sight tube and operator's eye form a sensor (a device which converts information into electrical signal).

information is also transmitted and processed in electrical form—however, through the transport of electrons. Sensors that are used in artificial systems must speak the same language as the devices with which they are interfaced. This language is electrical in its nature and a man-made sensor should be capable of responding with signals where information is carried by displacement of electrons, rather than ions.[1] Thus, it should be possible to connect a sensor to an electronic system through electrical wires, rather than through an electrochemical solution or a nerve fiber. Hence, in this book, we use a somewhat narrower definition of sensors, which may be phrased as

> A sensor is a device that receives a stimulus and responds with an electrical signal.

The term *stimulus* is used throughout this book and needs to be clearly understood. The stimulus is the quantity, property, or condition that is sensed and converted into electrical signal. Some texts (for instance, Ref. [2]) use a different term, *measurand*, which has the same meaning, however with the stress on quantitative characteristic of sensing.

The purpose of a sensor is to respond to some kind of an input physical property (stimulus) and to convert it into an electrical signal which is compatible with electronic circuits. We may say that a sensor is a translator of a generally nonelectrical value into an electrical value. When we say "electrical," we mean a signal which can be channeled, amplified, and modified by electronic devices. The sensor's output signal may be in the form of voltage, current, or charge. These may be further described in terms of amplitude, frequency, phase, or digital code. This set of characteristics is called the *output signal format*. Therefore, a sensor has input properties (of any kind) and electrical output properties.

[1] There is a very exciting field of the optical computing and communications where information is processed by a transport of photons. That field is beyond the scope of this book.

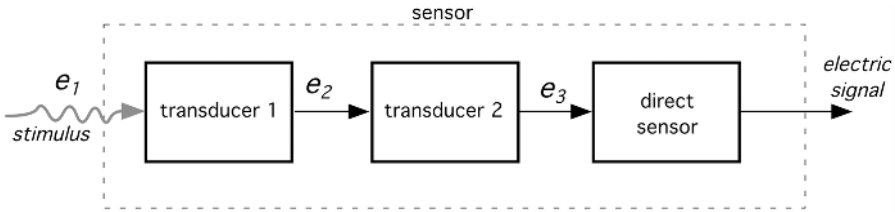


Fig. 1.2. A sensor may incorporate several transducers. e_1, e_2, and so on are various types of energy. Note that the last part is a direct sensor.

Any sensor is an energy converter. No matter what you try to measure, you always deal with energy transfer from the object of measurement to the sensor. The process of sensing is a particular case of information transfer, and any transmission of information requires transmission of energy. Of course, one should not be confused by an obvious fact that transmission of energy can flow both ways—it may be with a positive sign as well as with a negative sign; that is, energy can flow either from an object to the sensor or from the sensor to the object. A special case is when the energy is zero, and it also carries information about existence of that particular case. For example, a thermopile infrared radiation sensor will produce a positive voltage when the object is warmer than the sensor (infrared flux is flowing to the sensor) or the voltage is negative when the object is cooler than the sensor (infrared flux flows from the sensor to the object). When both the sensor and the object are at the same temperature, the flux is zero and the output voltage is zero. This carries a message that the temperatures are the same.

The term *sensor* should be distinguished from *transducer*. The latter is a converter of one type of energy into another, whereas the former converts any type of energy into *electrical*. An example of a transducer is a loudspeaker which converts an electrical signal into a variable magnetic field and, subsequently, into acoustic waves.[2] This is nothing to do with perception or sensing. Transducers may be used as *actuators* in various systems. An actuator may be described as opposite to a sensor—it converts electrical signal into generally nonelectrical energy. For example, an electric motor is an actuator—it converts electric energy into mechanical action.

Transducers may be parts of complex sensors (Fig. 1.2). For example, a chemical sensor may have a part which converts the energy of a chemical reaction into heat (transducer) and another part, a thermopile, which converts heat into an electrical signal. The combination of the two makes a chemical sensor—a device which produces an *electrical* signal in response to a chemical reaction. Note that in the above example, a chemical sensor is a complex sensor; it is comprised of a transducer and another sensor (heat). This suggests that many sensors incorporate at least one *direct*-type sensor and a number of transducers. The direct sensors are those that employ such physical effects that make a *direct energy conversion into electrical signal generation or modification*. Examples of such physical effects are photoeffect and Seebeck effect. These will be described in Chapter 3.

[2] It is interesting to note that a loudspeaker, when connected to an input of an amplifier, may function as a microphone. In that case, it becomes an acoustical sensor.

In summary, there are two types of sensors: *direct* and *complex*. A direct sensor converts a stimulus into an electrical signal or modifies an electrical signal by using an appropriate physical effect, whereas a complex sensor in addition needs one or more transducers of energy before a direct sensor can be employed to generate an electrical output.

A sensor does not function by itself; it is always a part of a larger system that may incorporate many other detectors, signal conditioners, signal processors, memory devices, data recorders, and actuators. The sensor's place in a device is either intrinsic or extrinsic. It may be positioned at the input of a device to perceive the outside effects and to signal the system about variations in the outside stimuli. Also, it may be an internal part of a device that monitors the devices' own state to cause the appropriate performance. A sensor is always a part of some kind of a data acquisition system. Often, such a system may be a part of a larger control system that includes various feedback mechanisms.

To illustrate the place of sensors in a larger system, Fig. 1.3 shows a block diagram of a data acquisition and control device. An object can be anything: a car, space ship, animal or human, liquid, or gas. Any material object may become a subject of some kind of a measurement. Data are collected from an object by a number of sensors. Some of them (2, 3, and 4) are positioned directly on or inside the object. Sensor 1 perceives the object without a physical contact and, therefore, is called a *noncontact* sensor. Examples of such a sensor is a radiation detector and a TV camera. Even if

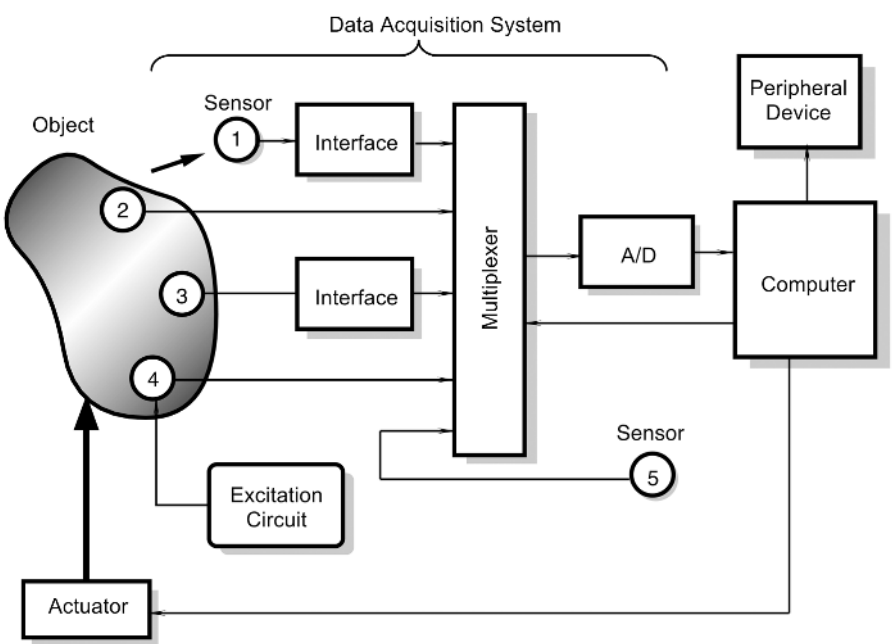


Fig. 1.3. Positions of sensors in a data acquisition system. Sensor 1 is noncontact, sensors 2 and 3 are passive, sensor 4 is active, and sensor 5 is internal to a data acquisition system.

we say "noncontact", we remember that energy transfer always occurs between any sensor and an object.

Sensor 5 serves a different purpose. It monitors internal conditions of a data acquisition system itself. Some sensors (1 and 3) cannot be directly connected to standard electronic circuits because of inappropriate output signal formats. They require the use of interface devices (signal conditioners). Sensors 1, 2, 3, and 5 are passive. They generate electric signals without energy consumption from the electronic circuits. Sensor 4 is active. It requires an operating signal, which is provided by an excitation circuit. This signal is modified by the sensor in accordance with the converted information. An example of an active sensor is a thermistor, which is a temperature-sensitive resistor. It may operate with a constant-current source, which is an excitation circuit. Depending on the complexity of the system, the total number of sensors may vary from as little as one (a home thermostat) to many thousands (a space shuttle).

Electrical signals from the sensors are fed into a multiplexer (MUX), which is a switch or a gate. Its function is to connect sensors one at a time to an analog-to-digital (A/D) converter if a sensor produces an analog signal, or directly to a computer if a sensor produces signals in a digital format. The computer controls a multiplexer and an A/D converter for the appropriate timing. Also, it may send control signals to the actuator, which acts on the object. Examples of actuators are an electric motor, a solenoid, a relay, and a pneumatic valve. The system contains some peripheral devices (for instance, a data recorder, a display, an alarm, etc.) and a number of components, which are not shown in the block diagram. These may be filters, sample-and-hold circuits, amplifiers, and so forth.

To illustrate how such a system works, let us consider a simple car-door monitoring arrangement. Every door in a car is supplied with a sensor which detects the door position (open or closed). In most cars, the sensor is a simple electric switch. Signals from all door sensors go to the car's internal microprocessor (no need for an A/D converter as all door signals are in a digital format: ones or zeros). The microprocessor identifies which door is open and sends an indicating signal to the peripheral devices (a dashboard display and an audible alarm). A car driver (the actuator) gets the message and acts on the object (closes the door).

An example of a more complex device is an anesthetic vapor delivery system. It is intended for controlling the level of anesthetic drugs delivered to a patient by means of inhalation during surgical procedures. The system employs several active and passive sensors. The vapor concentration of anesthetic agents (such as halothane, isoflurane, or enflurane) is selectively monitored by an active piezoelectric sensor, installed into a ventilation tube. Molecules of anesthetic vapors add mass to the oscillating crystal in the sensor and change its natural frequency, which is a measure of vapor concentration. Several other sensors monitor the concentration of CO_2, to distinguish exhale from inhale, and temperature and pressure, to compensate for additional variables. All of these data are multiplexed, digitized, and fed into the microprocessor, which calculates the actual vapor concentration. An anesthesiologist presets a desired delivery level and the processor adjusts the actuator (the valves) to maintain anesthetics at the correct concentration.

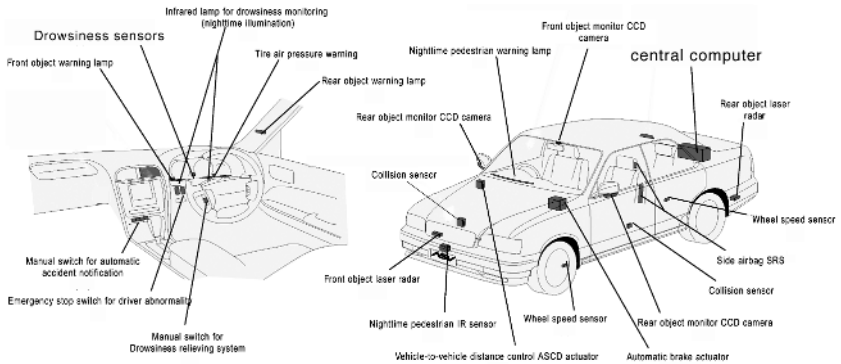

Drowsiness sensors
Infrared lamp for drowsiness monitoring (nighttime illumination)
Front object warning lamp
Tire air pressure warning
Rear object warning lamp
Manual switch for automatic accident notification
Emergency stop switch for driver abnormality
Manual switch for Drowsiness relieving system
central computer
Front object monitor CCD camera
Nighttime pedestrian warning lamp
Rear object laser radar
Rear object monitor CCD camera
Collision sensor
Wheel speed sensor
Side airbag SRS
Front object laser radar
Collision sensor
Nighttime pedestrian IR sensor
Wheel speed sensor
Rear object monitor CCD camera
Vehicle-to-vehicle distance control ASCD actuator
Automatic brake actuator

Fig. 1.4. Multiple sensors, actuators, and warning signals are parts of the Advanced Safety Vehicle. (Courtesy of Nissan Motor Company.)

Another example of a complex combination of various sensors, actuators, and indicating signals is shown in Fig. 1.4. It is an Advanced Safety Vehicle (ASV) that is being developed by Nissan. The system is aimed at increasing safety of a car. Among many others, it includes a drowsiness warning system and drowsiness relieving system. This may include the eyeball movement sensor and the driver head inclination detector. The microwave, ultrasonic, and infrared range measuring sensors are incorporated into the emergency braking advanced advisory system to illuminate the break lamps even before the driver brakes hard in an emergency, thus advising the driver of a following vehicle to take evasive action. The obstacle warning system includes both the radar and infrared (IR) detectors. The adaptive cruise control system works if the driver approaches too closely to a preceding vehicle: The speed is automatically reduced to maintain a suitable safety distance. The pedestrian monitoring system detects and alerts the driver to the presence of pedestrians at night as well as in vehicle blind spots. The lane control system helps in the event that the system detects and determines that incipient lane deviation is not the driver's intention. It issues a warning and automatically steers the vehicle, if necessary, to prevent it from leaving its lane.

In the following chapters, we concentrate on methods of sensing, physical principles of sensors operations, practical designs, and interface electronic circuits. Other essential parts of the control and monitoring systems, such as actuators, displays, data recorders, data transmitters, and others, are beyond the scope of this book and mentioned only briefly.

Generally, the sensor's input signals (stimuli) may have almost any conceivable physical or chemical nature (e.g., light flux, temperature, pressure, vibration, displacement, position, velocity, ion concentration, ...). The sensor's design may be of a general purpose. A special packaging and housing should be built to adapt it for a particular application. For instance, a micromachined piezoresistive pressure sensor may be housed into a watertight enclosure for the invasive measurement of aortic blood pressure through a catheter. The same sensor will be given an entirely different enclosure when it is intended for measuring blood pressure by a noninvasive

oscillometric method with an inflatable cuff. Some sensors are specifically designed to be very selective in a particular range of input stimulus and be quite immune to signals outside of the desirable limits. For instance, a motion detector for a security system should be sensitive to movement of humans and not responsive to movement of smaller animals, like dogs and cats.

1.2 Sensor Classification

Sensor classification schemes range from very simple to the complex. Depending on the classification purpose, different classification criteria may be selected. Here, we offer several practical ways to look at the sensors.

All sensors may be of two kinds: **passive** and **active**. A passive sensor does not need any additional energy source and directly generates an electric signal in response to an external stimulus; that is, the input stimulus energy is converted by the sensor into the output signal. The examples are a thermocouple, a photodiode, and a piezoelectric sensor. Most of passive sensors are direct sensors as we defined them earlier. The active sensors require external power for their operation, which is called an *excitation signal*. That signal is modified by the sensor to produce the output signal. The active sensors sometimes are called *parametric* because their own properties change in response to an external effect and these properties can be subsequently converted into electric signals. It can be stated that a sensor's parameter modulates the excitation signal and that modulation carries information of the measured value. For example, a thermistor is a temperature-sensitive resistor. It does not generate any electric signal, but by passing an electric current through it (excitation signal), its resistance can be measured by detecting variations in current and/or voltage across the thermistor. These variations (presented in ohms) directly relate to ttemperature through a known function. Another example of an active sensor is a resistive strain gauge in which electrical resistance relates to a strain. To measure the resistance of a sensor, electric current must be applied to it from an external power source.

Depending on the selected reference, sensors can be classified into **absolute** and **relative**. An *absolute* sensor detects a stimulus in reference to an absolute physical scale that is independent on the measurement conditions, whereas a *relative* sensor produces a signal that relates to some special case. An example of an absolute sensor is a thermistor: a temperature-sensitive resistor. Its electrical resistance directly relates to the absolute temperature scale of Kelvin. Another very popular temperature sensor—a thermocouple—is a relative sensor. It produces an electric voltage that is function of a temperature gradient across the thermocouple wires. Thus, a thermocouple output signal cannot be related to any particular temperature without referencing to a known baseline. Another example of the absolute and relative sensors is a pressure sensor. An absolute-pressure sensor produces signal in reference to vacuum—an absolute zero on a pressure scale. A relative-pressure sensor produces signal with respect to a selected baseline that is not zero pressure (e.g., to the atmospheric pressure).

Another way to look at a sensor is to consider all of its properties, such as what it measures (stimulus), what its specifications are, what physical phenomenon it is

sensitive to, what conversion mechanism is employed, what material it is fabricated from, and what its field of application is. Tables 1.1–1.6, adapted from Ref. [3], represent such a classification scheme, which is pretty much broad and representative. If we take for the illustration a surface acoustic-wave oscillator accelerometer, the table entries might be as follows:

Stimulus:	Acceleration
Specifications:	Sensitivity in frequency shift per gram of acceleration, short- and long-term stability in Hz per unit time, etc.
Detection means:	Mechanical
Conversion phenomenon:	Elastoelectric
Material:	Inorganic insulator
Field:	Automotive, marine, space, and scientific measurement

Table 1.1. Specifications

Sensitivity	Stimulus range (span)
Stability (short and long term)	Resolution
Accuracy	Selectivity
Speed of response	Environmental conditions
Overload characteristics	Linearity
Hysteresis	Dead band
Operating life	Output format
Cost, size, weight	Other

Table 1.2. Sensor Material

Inorganic	Organic
Conductor	Insulator
Semiconductor	Liquid, gas, or plasma
Biological substance	Other

Table 1.3. Detection Means Used in Sensors

Biological
Chemical
Electric, magnetic, or electromagnetic wave
Heat, temperature
Mechanical displacement or wave
Radioactivity, radiation
Other

Table 1.4. Conversion Phenomena

Physical	Chemical
Thermoelectric	Chemical transformation
Photoelectric	Physical transformation
Photomagnetic	Electrochemical process
Magnetoelectric	Spectroscopy
Electromagnetic	Other
Thermoelastic	Biological
Electroelastic	Biochemical transformation
Thermomagnetic	Physical transformation
Thermooptic	Effect on test organism
Photoelastic	Spectroscopy
Other	Other

Table 1.5. Field of Applications

Agriculture	Automotive
Civil engineering, construction	Domestic, appliances
Distribution, commerce, finance	Environment, meteorology, security
Energy, power	Information, telecommunication
Health, medicine	Marine
Manufacturing	Recreation, toys
Military	Space
Scientific measurement	Other
Transportation (excluding automotive)	

1.3 Units of Measurements

In this book, we use base units which have been established in The 14th General Conference on Weights and Measures (1971). The base measurement system is known as SI which stands for French *"Le Systéme International d'Unités"* (Table 1.7) [4]. All other physical quantities are derivatives of these base units. Some of them are listed in Table A.3.

Often, it is not convenient to use base or derivative units directly; in practice, quantities may be either too large or too small. For convenience in the engineering work, multiples and submultiples of the units are generally employed. They can be obtained by multiplying a unit by a factor from Table A.2. When pronounced, in all cases the first syllable is accented. For example, 1 ampere (A) may be multiplied by factor of 10^{-3} to obtain a smaller unit: 1 milliampere (mA), which is one-thousandth of an ampere.

Sometimes, two other systems of units are used. They are the Gaussian System and the British System, which in the United States its modification is called the *U.S. Customary System*. The United States is the only developed country in which SI

Table 1.6. Stimulus

Acoustic	Mechanical
Wave amplitude, phase, polarization	Position (linear, angular)
Spectrum	Acceleration
Wave velocity	Force
Other	Stress, pressure
Biological	Strain
Biomass (types, concentration, states)	Mass, density
Other	Moment, torque
Chemical	Speed of flow,rate of mass transport
Components (identities, concentration, states)	Shape, roughness, orientation
Other	Stiffness, compliance
Electric	Viscosity
Charge, current	Crystallinity, structural integrity
Potential, voltage	Other
Electric field (amplitude, phase, polarization, spectrum)	Radiation
	Type
Conductivity	Energy
Permitivity	Intensity
Other	Other
Magnetic	Thermal
Magnetic field (amplitude, phase, polarization, spectrum)	Temperature
	Flux
Magnetic flux	Specific heat
Permeability	Thermal conductivity
Other	Other
Optical	
Wave amplitude, phase, polarization, spectrum	
Wave velocity	
Refractive index	
Emissivity	
reflectivity, absorption	
Other	

still is not in common use. However, with the end of communism and the increase of world integration, international cooperation gains strong momentum. Hence, it is unavoidable that the United States will convert to SI[3] in the future, although maybe not in our lifetime. Still, in this book, we will generally use SI; however, for the convenience of the reader, the U.S. customary system units will be used in places where U.S. manufacturers employ them for sensor specifications. For the conversion to SI from other systems,[4] the reader may use Tables A.4. To make a conversion, a

[3] SI is often called the modernized metric system.

[4] Nomenclature, abbreviations, and spelling in the conversion tables are in accordance with "Standard practice for use of the International System of units (SI) (the Modernized Metric System)". Standard E380-91a. ©1991 ASTM, West Conshocken, PA.

Table 1.7. SI Basic Units

Quantity	Name	Symbol	Defined by (Year Established)
Length	Meter	m	The length of the path traveled by light in vacuum in 1/299,792,458 of a second. (1983)
Mass	Kilogram	kg	After a platinum–iridium prototype (1889)
Time	Second	s	The duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom (1967)
Electric current	Ampere	A	Force equal to 2×10^{-7} Nm of length exerted on two parallel conductors in vacuum when they carry the current (1946)
Thermodynamic temperature	Kelvin	K	The fraction 1/273.16 of the thermodynamic temperature of the triple point of water length(1967)
Amount of substance	Mole	mol	The amount of substance which contains as many elementary entities as there are atoms in 0.012 kg of carbon 12 (1971)
Luminous intensity	Candela	cd	Intensity in the perpendicular direction of a surface of 1/600,000 m^2 of a blackbody at temperature of freezing Pt under pressure of 101,325 Nm^2 (1967)
Plane angle	Radian	rad	(Supplemental unit)
Solid angle	Steradian	sr	(Supplemental unit)

non-SI value should be multiplied by a number given in the table. For instance, to convert an acceleration of 55 ft/s^2 to SI, it must to be multiplied by 0.3048:

$$55 \text{ft/s}^2 \times 0.3048 = 16.764 \text{ m/s}^2$$

Similarly, to convert an electric charge of 1.7 faraday, it must be multiplied by 9.65×10^{19}:

$$1.7 \text{ faraday} \times 9.65 \times 10^{19} = 1.64 \times 10^{20} \text{C}$$

The reader should consider the correct terminology of the physical and technical terms. For example, in the United States and many other countries, the electric potential difference is called "voltage," whereas in other countries, "electric tension" or simply "tension" is in common use. In this book, we use terminology that is traditional in the United States.

References

1. Thompson, S. *Control Systems: Engineering & Design*. Longman Scientific & Technical, Essex, UK, 1989.

2. Norton, H. N. *Handbook of Transducers*. Prentice-Hall, Englewood Cliffs, NJ, 1989.
3. White, R. W. A sensor classification scheme. In: *Microsensors*. IEEE Press, New York, 1991, pp. 3–5.
4. *The International System of Units (SI)*. B.N. Taylor, ed., NIST Special Publication 330, 2001.

2

Sensor Characteristics

"O, what men dare do! What men may do!
What men daily do, not knowing what they do."

—Shakespeare, "Much Ado About Nothing"

From the input to the output, a sensor may have several conversion steps before it produces an electrical signal. For instance, pressure inflicted on the fiber-optic sensor first results in strain in the fiber, which, in turn, causes deflection in its refractive index, which, in turn, results in an overall change in optical transmission and modulation of photon density. Finally, photon flux is detected and converted into electric current. In this chapter, we discuss the overall sensor characteristics, regardless of its physical nature or steps required to make a conversion. We regard a sensor as a "black box" where we are concerned only with relationships between its output signal and input stimulus.

2.1 Transfer Function

An *ideal* or *theoretical* output–stimulus relationship exists for every sensor. If the sensor is ideally designed and fabricated with ideal materials by ideal workers using ideal tools, the output of such a sensor would always represent the *true* value of the stimulus. The ideal function may be stated in the form of a table of values, a graph, or a mathematical equation. An ideal (theoretical) output–stimulus relationship is characterized by the so-called *transfer function*. This function establishes dependence between the electrical signal S produced by the sensor and the stimulus $s : S = f(s)$. That function may be a simple linear connection or a nonlinear dependence, (e.g., logarithmic, exponential, or power function). In many cases, the relationship is unidimensional (i.e., the output versus one input stimulus). A unidimensional linear relationship is represented by the equation

$$S = a + bs, \tag{2.1}$$

where a is the intercept (i.e., the output signal at zero input signal) and b is the slope, which is sometimes called *sensitivity*. S is one of the characteristics of the output electric signal used by the data acquisition devices as the sensor's output. It may be amplitude, frequency, or phase, depending on the sensor properties.

Logarithmic function:

$$S = a + b \ln s. \tag{2.2}$$

Exponential function:

$$S = ae^{ks}. \tag{2.3}$$

Power function:

$$S = a_0 + a_1 s^k, \tag{2.4}$$

where k is a constant number.

A sensor may have such a transfer function that none of the above approximations fits sufficiently well. In that case, a higher-order polynomial approximation is often employed.

For a nonlinear transfer function, the sensitivity b is not a fixed number as for the linear relationship [Eq. (2.1)]. At any particular input value, s_0, it can be defined as

$$b = \frac{dS(s_0)}{ds}. \tag{2.5}$$

In many cases, a nonlinear sensor may be considered linear over a limited range. Over the extended range, a nonlinear transfer function may be modeled by several straight lines. This is called a piecewise approximation. To determine whether a function can be represented by a linear model, the incremental variables are introduced for the input while observing the output. A difference between the actual response and a liner model is compared with the specified accuracy limits (see 2.4).

A transfer function may have more than one dimension when the sensor's output is influenced by more than one input stimuli. An example is the transfer function of a thermal radiation (infrared) sensor. The function[1] connects two temperatures (T_b, the absolute temperature of an object of measurement, and T_s, the absolute temperature of the sensor's surface) and the output voltage V:

$$V = G(T_b^4 - T_s^4), \tag{2.6}$$

where G is a constant. Clearly, the relationship between the object's temperature and the output voltage (transfer function) is not only nonlinear (the fourth-order parabola) but also depends on the sensor's surface temperature. To determine the sensitivity of the sensor with respect to the object's temperature, a partial derivative will be calculated as

$$b = \frac{\partial V}{\partial T_b} = 4GT_b^3. \tag{2.7}$$

The graphical representation of a two-dimensional transfer function of Eq. (2.6) is shown in Fig. 2.1. It can be seen that each value of the output voltage can be uniquely

[1] This function is generally known as the Stefan–Boltzmann law.

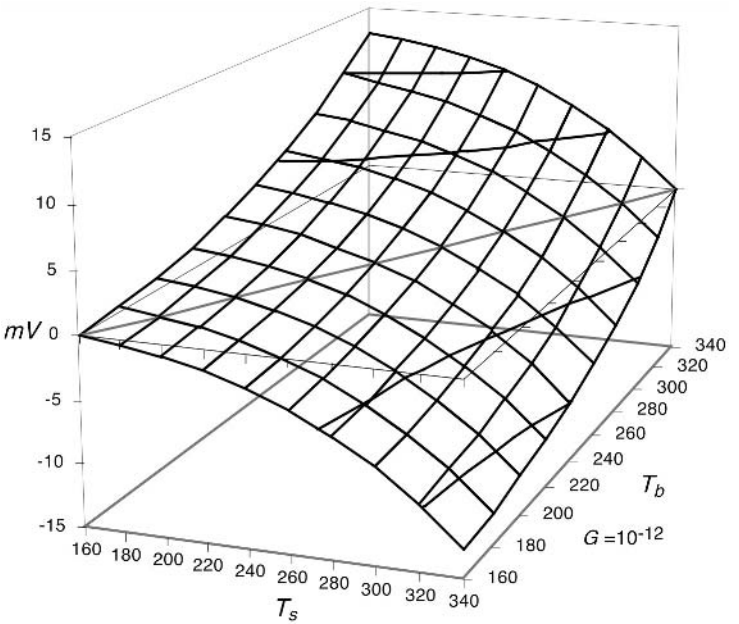


Fig. 2.1. Two-dimensional transfer function of a thermal radiation sensor.

determined from two input temperatures. It should be noted that a transfer function represents the input-to-output relationship. However, when a sensor is used for measuring or detecting a stimulus, an inversed function (output-to-input) needs to be employed. When a transfer function is linear, the inversed function is very easy to compute. When it is nonlinear the task is more complex, and in many cases, the analytical solution may not lend itself to reasonably simple data processing. In these cases, an approximation technique often is the solution.

2.2 Span (Full-Scale Input)

A dynamic range of stimuli which may be converted by a sensor is called a *span* or an *input full scale* (FS). It represents the highest possible input value that can be applied to the sensor without causing an unacceptably large inaccuracy. For the sensors with a very broad and nonlinear response characteristic, a dynamic range of the input stimuli is often expressed in decibels, which is a logarithmic measure of ratios of either power or force (voltage). It should be emphasized that decibels do not measure absolute values, but a ratio of values only. A decibel scale represents signal magnitudes by much smaller numbers, which, in many cases, is far more convenient. Being a nonlinear scale, it may represent low-level signals with high resolution while compressing the high-level numbers. In other words, the logarithmic scale for small objects works as a microscope, and for the large objects, it works as a telescope. By

Table 2.1. Relationship Among Power, Force (Voltage, Current), and Decibels

Power ratio	1.023	1.26	10.0	100	10^3	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}
Force ratio	1.012	1.12	3.16	10.0	31.6	100	316	10^3	3162	10^4	3×10^4	10^5
Decibels	0.1	1.0	10.0	20.0	30.0	40.0	50.0	60.0	70.0	80.0	90.0	100.0

definition, decibels are equal to 10 times the log of the ratio of powers (Table 2.1):

$$1 \text{ dB} = 10 \log \frac{P_2}{P_1}. \tag{2.8}$$

In a similar manner, decibels are equal to 20 times the log of the force, current, or voltage:

$$1 \text{ dB} = 20 \log \frac{S_2}{S_1}. \tag{2.9}$$

2.3 Full-Scale Output

Full-scale output (FSO) is the algebraic difference between the electrical output signals measured with maximum input stimulus and the lowest input stimulus applied. This must include all deviations from the ideal transfer function. For instance, the FSO output in Fig. 2.2A is represented by S_{FS}.

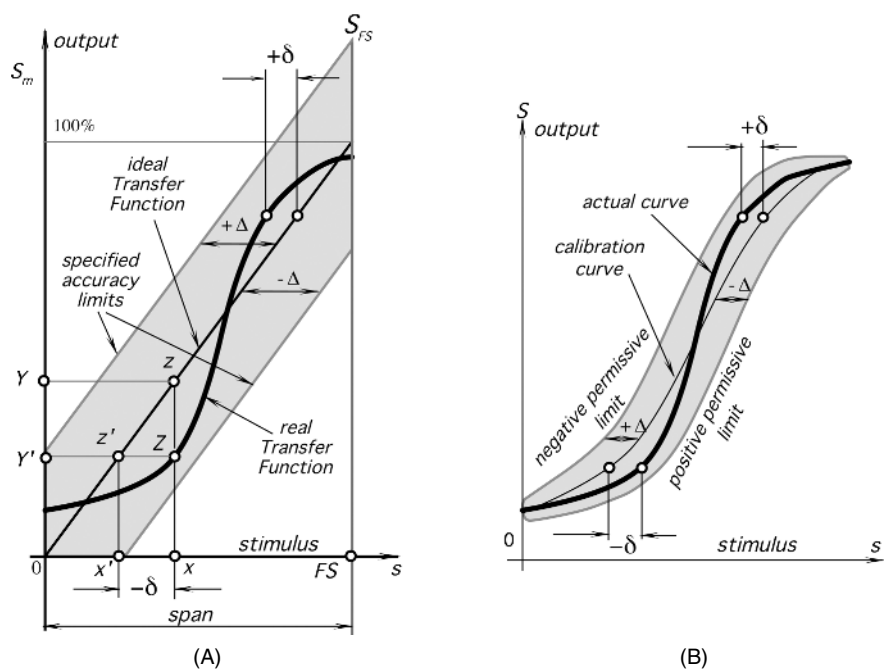


Fig. 2.2. Transfer function (A) and accuracy limits (B). Error is specified in terms of input value.

2.4 Accuracy

A very important characteristic of a sensor is *accuracy* which really means *inaccuracy*. Inaccuracy is measured as a highest deviation of a value represented by the sensor from the ideal or true value at its input. The true value is attributed to the object of measurement and accepted as having a specified uncertainty (see 2.20.)

The deviation can be described as a difference between the value which is computed from the output voltage and the actual input value. For example, a linear displacement sensor ideally should generate 1 mV per 1-mm displacement; that is, its transfer function is linear with a slope (sensitivity) $b = 1$ mV/mm. However, in the experiment, a displacement of $s = 10$ mm produced an output of $S = 10.5$ mV. Converting this number into the displacement value by using the inversed transfer function ($1/b = 1$ mm/mV), we would calculate that the displacement was $s_x = S/b = 10.5$ mm; that is $s_x - s = 0.5$ mm more than the actual. This extra 0.5 mm is an erroneous deviation in the measurement, or error. Therefore, in a 10-mm range, the sensor's absolute inaccuracy is 0.5 mm, or in the relative terms, inaccuracy is $(0.5\text{mm}/10\text{mm}) \times 100\% = 5\%$. If we repeat this experiment over and over again without any random error and every time we observe an error of 0.5 mm, we may say that the sensor has a *systematic* inaccuracy of 0.5 mm over a 10-mm span. Naturally, a random component is always present, so the systematic error may be represented as an average or mean value of multiple errors.

Figure 2.2A shows an ideal or theoretical transfer function. In the real world, any sensor performs with some kind of imperfection. A possible *real* transfer function is represented by a thick line, which generally may be neither linear nor monotonic. A real function rarely coincides with the ideal. Because of material variations, workmanship, design errors, manufacturing tolerances, and other limitations, it is possible to have a large family of real transfer functions, even when sensors are tested under identical conditions. However, all runs of the real transfer functions must fall within the limits of a specified accuracy. These permissive limits differ from the ideal transfer function line by $\pm\Delta$. The real functions deviate from the ideal by $\pm\delta$, where $\delta \leq \Delta$. For example, let us consider a stimulus having value x. Ideally, we would expect this value to correspond to point z on the transfer function, resulting in the output value Y. Instead, the real function will respond at point Z, producing output value Y'. This output value corresponds to point z' on the ideal transfer function, which, in turn, relates to a "would-be" input stimulus x' whose value is smaller than x. Thus, in this example, imperfection in the sensor's transfer function leads to a measurement error of $-\delta$.

The accuracy rating includes a combined effect of part-to-part variations, a hysteresis, a dead band, calibration, and repeatability errors (see later subsections). The specified accuracy limits generally are used in the worst-case analysis to determine the worst possible performance of the system. Figure 2.2B shows that $\pm\Delta$ may more closely follow the real transfer function, meaning better tolerances of the sensor's accuracy. This can be accomplished by a multiple-point calibration. Thus, the specified accuracy limits are established not around the theoretical (ideal) transfer function, but around the calibration curve, which is determined during the actual calibration procedure. Then, the permissive limits become narrower, as they do not embrace

part-to-part variations between the sensors and are geared specifically to the calibrated unit. Clearly, this method allows more accurate sensing; however, in some applications, it may be prohibitive because of a higher cost.

The inaccuracy rating may be represented in a number of forms:

1. Directly in terms of measured value (Δ)
2. In percent of input span (full scale)
3. In terms of output signal

For example, a piezoresistive pressure sensor has a 100-kPa input full scale and a 10Ω full-scale output. Its inaccuracy may be specified as ±0.5%, ±500 Pa, or ±0.05Ω.

In modern sensors, specification of accuracy often is replaced by a more comprehensive value of *uncertainty* (see Section 2.20) because uncertainty is comprised of all distorting effects both systematic and random and is not limited to the inaccuracy of a transfer function.

2.5 Calibration

If the sensor's manufacturer's tolerances and tolerances of the interface (signal conditioning) circuit are broader than the required system accuracy, a calibration is required. For example, we need to measure temperature with an accuracy ±0.5°C; however, an available sensor is rated as having an accuracy of ±1°C. Does it mean that the sensor can not be used? No, it can, but that particular sensor needs to be calibrated; that is, its individual transfer function needs to be found during calibration. Calibration means the determination of specific variables that describe the overall transfer function. Overall means of the entire circuit, including the sensor, the interface circuit, and the A/D converter. The mathematical model of the transfer function should be known before calibration. If the model is linear [Eq. (2.1)], then the calibration should determine variables a and b; if it is exponential [Eq. (2.3)], variables a and k should be determined; and so on. Let us consider a simple linear transfer function. Because a minimum of two points are required to define a straight line, at least a two-point calibration is required. For example, if one uses a forward-biased semiconductor p-n junction for temperature measurement, with a high degree of accuracy its transfer function (temperature is the input and voltage is the output) can be considered linear:

$$v = a + bt. \tag{2.10}$$

To determine constants a and b, such a sensor should be subjected to two temperatures (t_1 and t_2) and two corresponding output voltages (v_1 and v_2) will be registered. Then, after substituting these values into Eq. (2.10), we arrive at

$$\begin{aligned} v_1 &= a + bt_1, \\ v_2 &= a + bt_2, \end{aligned} \tag{2.11}$$

and the constants are computed as

$$b = \frac{v_1 - v_2}{t_1 - t_2} \quad \text{and} \quad a = v_1 - bt_1. \tag{2.12}$$

To compute the temperature from the output voltage, a measured voltage is inserted into an inversed equation

$$t = \frac{v - a}{b}. \tag{2.13}$$

In some fortunate cases, one of the constants may be specified with a sufficient accuracy so that no calibration of that particular constant may be needed. In the same p-n-junction temperature sensor, the slope b is usually a very consistent value for a given lot and type of semiconductor. For example, a value of $b = -0.002268$ V/°C was determined to be consistent for a selected type of the diode, then a single-point calibration is needed to find out a as $a = v_1 + 0.002268t_1$.

For nonlinear functions, more than two points may be required, depending on a mathematical model of the transfer function. Any transfer function may be modeled by a polynomial, and depending on required accuracy, the number of the calibration points should be selected. Because calibration may be a slow process, to reduce production cost in manufacturing, it is very important to minimize the number of calibration points.

Another way to calibrate a nonlinear transfer function is to use a piecewise approximation. As was mentioned earlier, any section of a curvature, when sufficiently small, can be considered linear and modeled by Eq. (2.1). Then, a curvature will be described by a family of linear lines where each has its own constants a and b. During the measurement, one should determine where on the curve a particular output voltage S is situated and select the appropriate set of constants a and b to compute the value of a corresponding stimulus s from an equation identical to Eq. (2.13).

To calibrate sensors, it is essential to have and properly maintain precision and accurate physical standards of the appropriate stimuli. For example, to calibrate contact-temperature sensors, either a temperature-controlled water bath or a "dry-well" cavity is required. To calibrate the infrared sensors, a blackbody cavity would be needed. To calibrate a hygrometer, a series of saturated salt solutions are required to sustain a constant relative humidity in a closed container, and so on. It should be clearly understood that the sensing system accuracy is directly attached to the accuracy of the calibrator. An uncertainty of the calibrating standard must be included in the statement on the overall uncertainty, as explained in 2.20.

2.6 Calibration Error

The *calibration error* is inaccuracy permitted by a manufacturer when a sensor is calibrated in the factory. This error is of a systematic nature, meaning that it is added to all possible real transfer functions. It shifts the accuracy of transduction for each stimulus point by a constant. This error is not necessarily uniform over the range and may change depending on the type of error in the calibration. For example, let us consider a two-point calibration of a real linear transfer function (thick line in Fig. 2.3). To determine the slope and the intercept of the function, two stimuli, s_1 and s_2, are applied to the sensor. The sensor responds with two corresponding output signals A_1 and A_2. The first response was measured absolutely accurately, however,

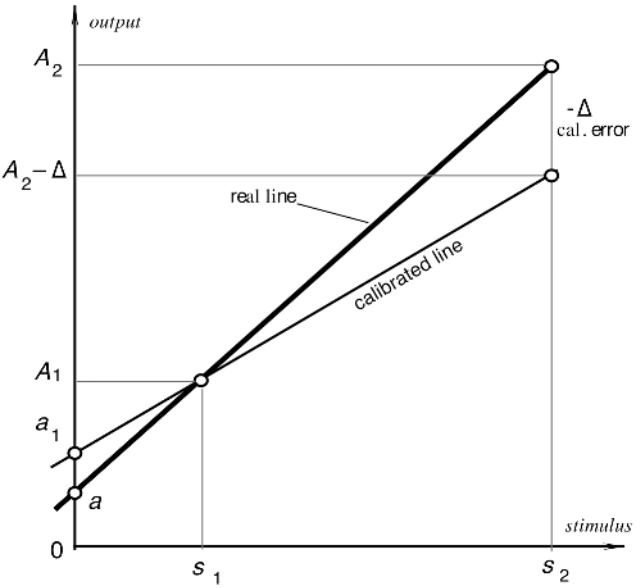


Fig. 2.3. Calibration error.

the higher signal was measured with error $-\Delta$. This results in errors in the slope and intercept calculation. A new intercept, a_1, will differ from the real intercept, a, by

$$\delta_a = a_1 - a = \frac{\Delta}{s_2 - s_1}, \tag{2.14}$$

and the slope will be calculated with error:

$$\delta_b = -\frac{\Delta}{s_2 - s_1}, \tag{2.15}$$

2.7 Hysteresis

A *hysteresis error* is a deviation of the sensor's output at a specified point of the input signal when it is approached from the opposite directions (Fig. 2.4). For example, a displacement sensor when the object moves from left to right at a certain point produces a voltage which differs by 20 mV from that when the object moves from right to left. If the sensitivity of the sensor is 10 mV/mm, the hysteresis error in terms of displacement units is 2 mm. Typical causes for hysteresis are friction and structural changes in the materials.

2.8 Nonlinearity

Nonlinearity error is specified for sensors whose transfer function may be approximated by a straight line [Eq. (2.1)]. A nonlinearity is a maximum deviation (L) of a real transfer function from the approximation straight line. The term "linearity" actually

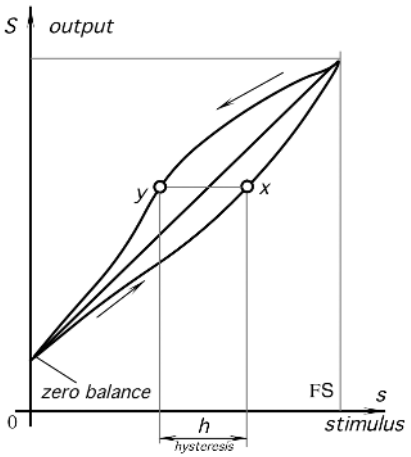


Fig. 2.4. Transfer function with hysteresis.

means "nonlinearity." When more than one calibration run is made, the worst linearity seen during any one calibration cycle should be stated. Usually, it is specified either in percent of span or in terms of measured value (e.g, in kPa or °C). "Linearity," when not accompanied by a statement explaining what sort of straight line it is referring to, is meaningless. There are several ways to specify a nonlinearity, depending how the line is superimposed on the transfer function. One way is to use *terminal* points (Fig. 2.5A); that is, to determine output values at the smallest and highest stimulus values and to draw a straight line through these two points (line 1). Here, near the terminal points, the nonlinearity error is the smallest and it is higher somewhere in between.

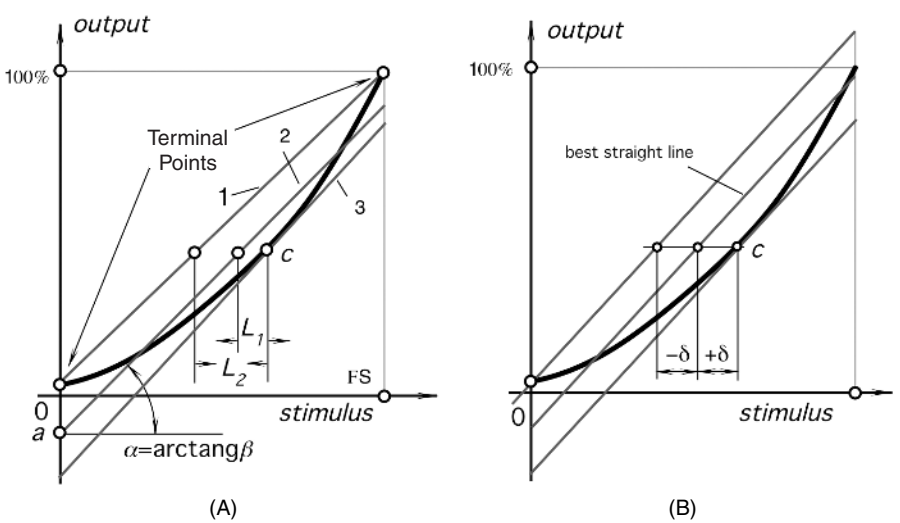


Fig. 2.5. Linear approximations of a nonlinear transfer function (A) and independent linearity (B).

Another way to define the approximation line is to use a method of *least squares* (line 2 in Fig. 2.5A). This can be done in the following manner. Measure several (n) output values S at input values s over a substantially broad range, preferably over an entire full scale. Use the following formulas for linear regression to determine intercept a and slope b of the best-fit straight line:

$$a = \frac{\sum S \sum s^2 - \sum s \sum sS}{n \sum s^2 - (\sum s)^2}, \qquad b = \frac{n \sum sS - \sum s \sum S}{n \sum s^2 - (\sum s)^2}, \tag{2.16}$$

where $\sum$ is the summation of n numbers.

In some applications, a higher accuracy may be desirable in a particular narrower section of the input range. For instance, a medical thermometer should have the best accuracy in a fever definition region which is between 37°C and 38°C. It may have a somewhat lower accuracy beyond these limits. Usually, such a sensor is calibrated in the region where the highest accuracy is desirable. Then, the approximation line may be drawn through the calibration point c (line 3 in Fig. 2.5A). As a result, nonlinearity has the smallest value near the calibration point and it increases toward the ends of the span. In this method, the line is often determined as tangent to the transfer function in point c. If the actual transfer function is known, the slope of the line can be found from Eq. (2.5).

Independent linearity is referred to as the so-called "best straight line" (Fig. 2.5B), which is a line midway between two parallel straight lines closest together and enveloping all output values on a real transfer function.

Depending on the specification method, approximation lines may have different intercepts and slopes. Therefore, nonlinearity measures may differ quite substantially from one another. A user should be aware that manufacturers often publish the smallest possible number to specify nonlinearity, without defining what method was used.

2.9 Saturation

Every sensor has its operating limits. Even if it is considered linear, at some levels of the input stimuli, its output signal no longer will be responsive. A further increase in stimulus does not produce a desirable output. It is said that the sensor exhibits a span-end nonlinearity or saturation (Fig. 2.6).

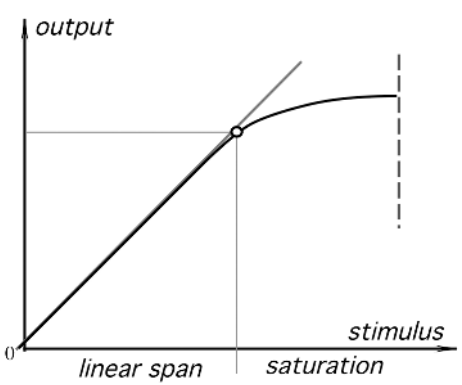


Fig. 2.6. Transfer function with saturation.

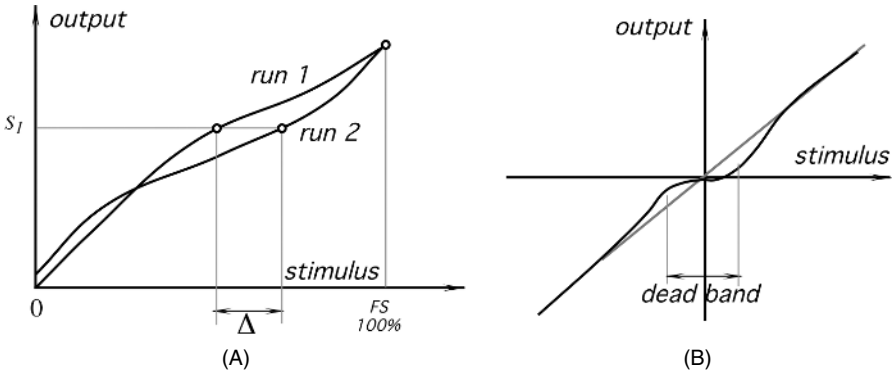


Fig. 2.7. (A) The repeatability error. The same output signal S_1 corresponds to two different input signals. (B) The dead-band zone in a transfer function.

2.10 Repeatability

A *repeatability* (reproducibility) error is caused by the inability of a sensor to represent the same value under identical conditions. It is expressed as the maximum difference between output readings as determined by two calibrating cycles (Fig. 2.7A), unless otherwise specified. It is usually represented as % of FS:

$$\delta_r = \frac{\Delta}{\text{FS}} \times 100\%. \tag{2.17}$$

Possible sources of the repeatability error may be thermal noise, buildup charge, material plasticity, and so forth.

2.11 Dead Band

The *dead band* is the insensitivity of a sensor in a specific range of input signals (Fig. 2.7B). In that range, the output may remain near a certain value (often zero) over an entire dead-band zone.

2.12 Resolution

Resolution describes the smallest increments of stimulus which can be sensed. When a stimulus continuously varies over the range, the output signals of some sensors will not be perfectly smooth, even under the no-noise conditions. The output may change in small steps. This is typical for potentiometric transducers, occupancy infrared detectors with grid masks, and other sensors where the output signal change is enabled only upon a certain degree of stimulus variation. In addition, any signal converted into a digital format is broken into small steps, where a number is assigned to each step. The magnitude of the input variation which results in the output smallest step is specified as resolution under specified conditions (if any). For instance, for the occupancy detector, the resolution may be specified as follows: “resolution—minimum

equidistant displacement of the object for 20 cm at 5 m distance." For wire-wound potentiometric angular sensors, resolution may be specified as "a minimum angle of 0.5°." Sometimes, it may be specified as percent of full scale (FS). For instance, for the angular sensor having 270° FS, the 0.5° resolution may be specified as 0.181% of FS. It should be noted that the step size may vary over the range, hence, the resolution may be specified as typical, average, or "worst." The resolution of digital output format sensors is given by the number of bits in the data word. For instance, the resolution may be specified as "8-bit resolution." To make sense, this statement must be accomplished with either the FS value or the value of LSB (least significant bit). When there are no measurable steps in the output signal, it is said that the sensor has *continuous* or *infinitesimal* resolution (sometimes erroneously referred to as "infinite resolution").

2.13 Special Properties

Special input properties may be needed to specify for some sensors. For instance, light detectors are sensitive within a limited optical bandwidth. Therefore, it is appropriate to specify a spectral response for them.

2.14 Output Impedance

The *output impedance* Z_{out} is important to know to better interface a sensor with the electronic circuit. This impedance is connected either in parallel with the input impedance Z_{in} of the circuit (voltage connection) or in series (current connection). Figure 2.8 shows these two connections. The output and input impedances generally should be represented in a complex form, as they may include active and reactive components. To minimize the output signal distortions, a current generating sensor (B) should have an output impedance as high as possible and the circuit's input impedance should be low. For the voltage connection (A), a sensor is preferable with lower Z_{out} and the circuit should have Z_{in} as high as practical.

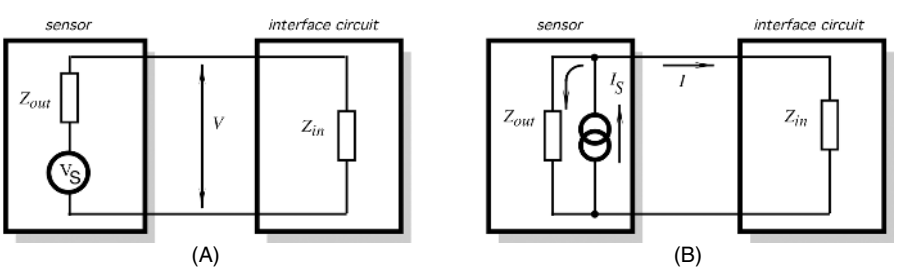


Fig. 2.8. Sensor connection to an interface circuit: (A) sensor has voltage output; (B) sensor has current output.

2.15 Excitation

Excitation is the electrical signal needed for the active sensor operation. Excitation is specified as a range of voltage and/or current. For some sensors, the frequency of the excitation signal and its stability must also be specified. Variations in the excitation may alter the sensor transfer function and cause output errors.

An example of excitation signal specification is as follows:

Maximum current through a thermistor
in still air 50 μA
in water 200 μA

2.16 Dynamic Characteristics

Under static conditions, a sensor is fully described by its transfer function, span, calibration, and so forth. However, when an input stimulus varies, a sensor response generally does not follow with perfect fidelity. The reason is that both the sensor and its coupling with the source of stimulus cannot always respond instantly. In other words, a sensor may be characterized with a *time*-dependent characteristic, which is called a *dynamic characteristic*. If a sensor does not respond instantly, it may indicate values of stimuli which are somewhat different from the real; that is, the sensor responds with a *dynamic error*. A difference between static and dynamic errors is that the latter is always time dependent. If a sensor is a part of a control system which has its own dynamic characteristics, the combination may cause, at best, a delay in representing a true value of a stimulus or, at worst, cause oscillations.

The *warm-up time* is the time between applying electric power to the sensor or excitation signal and the moment when the sensor can operate within its specified accuracy. Many sensors have a negligibly short warm-up time. However, some detectors, especially those that operate in a thermally controlled environment (a thermostat) may require seconds and minutes of warm-up time before they are fully operational within the specified accuracy limits.

In a control system theory, it is common to describe the input–output relationship through a constant-coefficient linear differential equation. Then, the sensor's dynamic (time-dependent) characteristics can be studied by evaluating such an equation. Depending on the sensor design, the differential equation can be of several *orders*.

A *zero-order* sensor is characterized by the relationship which, for a linear transfer function, is a modified Eq. (2.1) where the input and output are functions of time t:

$$S(t) = a + bs(t). \tag{2.18}$$

The value a is called an offset and b is called static sensitivity. Equation (2.18) requires that the sensor does not incorporate any energy storage device, like a capacitor or mass. A zero-order sensor responds instantaneously. In other words, such a sensor does not need any dynamic characteristics.

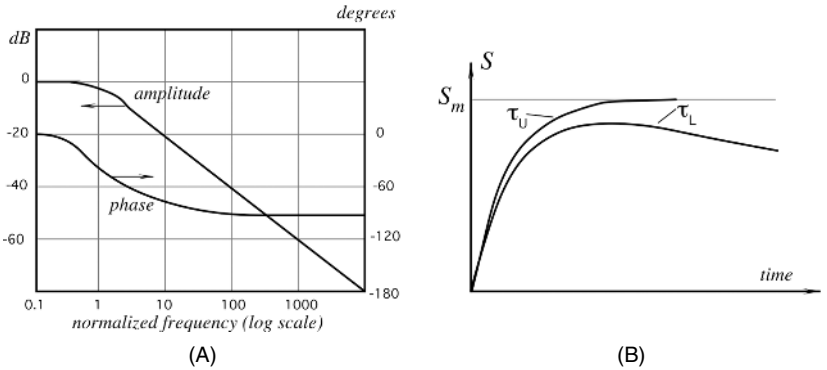


Fig. 2.9. Frequency characteristic (A) and response of a first-order sensor (B) with limited upper and lower cutoff frequencies. τ_u and τ_L are corresponding time constants.

A *first-order* differential equation describes a sensor that incorporates one energy storage component. The relationship between the input $s(t)$ and output $S(t)$ is the differential equation

$$b_1 \frac{dS(t)}{dt} + b_0 S(t) = s(t). \tag{2.19}$$

A typical example of a first-order sensor is a temperature sensor for which the energy storage is thermal capacity. The first-order sensors may be specified by a manufacturer in various ways. Typical is a *frequency response*, which specifies how fast a first-order sensor can react to a change in the input stimulus. The frequency response is expressed in hertz or rads per second to specify the relative reduction in the output signal at a certain frequency (Fig. 2.9A). A commonly used reduction number (frequency limit) is -3 dB. It shows at what frequency the output voltage (or current) drops by about 30%. The frequency response limit f_u is often called the upper cutoff frequency, as it is considered the highest frequency a sensor can process.

The frequency response directly relates to a *speed response*, which is defined in units of input stimulus per unit of time. Which response, frequency or speed, to specify in any particular case depends on the sensor type, its application, and the preference of a designer.

Another way to specify speed response is by time, which is required by the sensor to reach 90% of a steady-state or maximum level upon exposure to a step stimulus. For the first-order response, it is very convenient to use a so-called *time constant*. The time constant, τ, is a measure of the sensor's inertia. In electrical terms, it is equal to the product of electrical capacitance and resistance: $\tau = CR$. In thermal terms, thermal capacity and thermal resistances should be used instead. Practically, the time constant can be easily measured. A first-order system response is

$$S = S_m(1 - e^{-t/\tau}), \tag{2.20}$$

where S_m is steady-state output, t is time, and e is the base of natural logarithm.

Substituting $t = \tau$, we get

$$\frac{S}{S_m} = 1 - \frac{1}{e} = 0.6321. \tag{2.21}$$

In other words, after an elapse of time equal to one time constant, the response reaches about 63% of its steady-state level. Similarly, it can be shown that after two time constants, the height will be 86.5% and after three time constants it will be 95%.

The *cutoff frequency* indicates the lowest or highest frequency of stimulus that the sensor can process. The upper cutoff frequency shows how fast the sensor reacts; the lower cutoff frequency shows how slow the sensor can process changing stimuli. Figure 2.9B depicts the sensor's response when both the upper and lower cutoff frequencies are limited. As a rule of thumb, a simple formula can be used to establish a connection between the cutoff frequency, f_c (either upper and lower), and time constant in a first-order sensor:

$$f_c \approx \frac{0.159}{\tau}, \tag{2.22}$$

The *phase shift* at a specific frequency defines how the output signal lags behind in representing the stimulus change (Fig. 2.9A). The shift is measured in angular degrees or rads and is usually specified for a sensor that processes periodic signals. If a sensor is a part of a feedback control system, it is very important to know its phase characteristic. Phase lag reduces the phase margin of the system and may result in overall instability.

A *second-order* differential equation describes a sensor that incorporates two energy storage components. The relationship between the input $s(t)$ and output $S(t)$ is the differential equation

$$b_2 \frac{d^2 S(t)}{dt^2} + b_1 \frac{dS(t)}{dt} + b_0 S(t) = s(t). \tag{2.23}$$

An example of a second-order sensor is an accelerometer that incorporates a mass and a spring.

A second-order response is specific for a sensor that responds with a periodic signal. Such a periodic response may be very brief and we say that the sensor is damped, or it may be of a prolonged time and even may oscillate continuously. Naturally, for a sensor, such a continuous oscillation is a malfunction and must be avoided. Any second-order sensor may be characterized by a *resonant (natural) frequency*, which is a number expressed in hertz or rads per second. The natural frequency shows where the sensor's output signal increases considerably. Many sensors behave as if a dynamic sensor's output conforms to the standard curve of a second-order response; the manufacturer will state the natural frequency and the damping ratio of the sensor. The resonant frequency may be related to mechanical, thermal, or electrical properties of the detector. Generally, the operating frequency range for the sensor should be selected well below (at least 60%) or above the resonant frequency. However, in some sensors, the resonant frequency is the operating point. For instance, in glass-breakage detectors (used in security systems), the resonant makes the sensor selectively sensitive to a narrow bandwidth, which is specific for the acoustic spectrum produced by shattered glass.

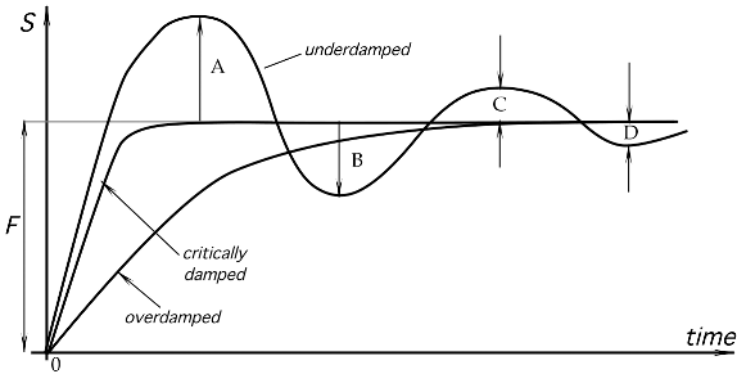


Fig. 2.10. Responses of sensors with different damping characteristics.

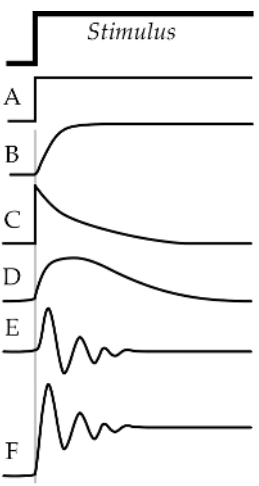


Fig. 2.11. Types of response: (A) unlimited upper and lower frequencies; (B) first-order limited upper cutoff frequency; (C) first-order limited lower cutoff frequency; (D) first-order limited both upper and lower cutoff frequencies; (E) narrow bandwidth response (resonant); (F) wide bandwidth with resonant.

Damping is the progressive reduction or suppression of the oscillation in the sensor having higher than a first-order response. When the sensor's response is as fast as possible without overshoot, the response is said to be critically damped (Fig. 2.10). An underdamped response is when the overshoot occurs and the overdamped response is slower than the critical response. The damping ratio is a number expressing the quotient of the actual damping of a second-order linear transducer by its critical damping.

For an oscillating response, as shown in Fig. 2.10, a *damping factor* is a measure of damping, expressed (without sign) as the quotient of the greater by the lesser of a pair of consecutive swings in opposite directions of the output signal, about an ultimately steady-state value. Hence, the damping factor can be measured as

$$\text{Damping factor} = \frac{F}{A} = \frac{A}{B} = \frac{B}{C} = \text{etc.} \tag{2.24}$$

2.17 Environmental Factors

Storage conditions are nonoperating environmental limits to which a sensor may be subjected during a specified period without permanently altering its performance under normal operating conditions. Usually, storage conditions include the highest and the lowest storage temperatures and maximum relative humidities at these temperatures. The word "noncondensing" may be added to the relative humidity number. Depending on the sensor's nature, some specific limitation for the storage may need to be considered (e.g., maximum pressure, presence of some gases or contaminating fumes, etc.).

Short- and long-term stabilities (drift) are parts of the accuracy specification. The short-term stability is manifested as changes in the sensor's performance within minutes, hours, or even days. The sensor's output signal may increase or decrease, which, in other terms, may be described as ultralow-frequency noise. The long-term stability may be related to *aging* of the sensor materials, which is an irreversible change in the material's electrical, mechanical, chemical, or thermal properties; that is, the long-term drift is usually unidirectional. It happens over a relatively long time span, such as months and years. Long-term stability is one of the most important for sensors used for precision measurements. Aging depends heavily on environmental storage and operating conditions, how well the sensor components are isolated from the environment, and what materials are used for their fabrication. The aging phenomenon is typical for sensors having organic components and, in general, is not an issue for a sensor made with only nonorganic materials. For instance, glass-coated metal-oxide thermistors exhibit much greater long-term stability compared to epoxy-coated thermistors. A powerful way to improve long-term stability is to preage the component at extreme conditions. The extreme conditions may be cycled from the lowest to the highest. For instance, a sensor may be periodically swung from freezing to hot temperatures. Such accelerated aging not only enhances the stability of the sensor's characteristics but also improves the reliability (see Section 2.18), as the preaging process reveals many hidden defects. For instance, epoxy-coated thermistors may be greatly improved if they are maintained at $+150°C$ for 1 month before they are calibrated and installed in a product.

Environmental conditions to which a sensor is subjected do not include variables which the sensor measures. For instance, an air-pressure sensor usually is subjected not just to air pressure but to other influences as well, such as the temperatures of air and surrounding components, humidity, vibration, ionizing radiation, electromagnetic fields, gravitational forces, and so forth. All of these factors may and usually do affect the sensor's performance. Both static and dynamic variations in these conditions should be considered. Some environmental conditions are usually of a multiplicative nature; that is, they alter a transfer function of the sensor (e.g., changing its gain). One example is the resistive strain gauge, whose sensitivity increases with temperature.

Environmental stability is quite broad and usually a very important requirement. Both the sensor designer and the application engineer should consider all possible external factors which may affect the sensor's performance. A piezoelectric accelerometer may generate spurious signals if affected by a sudden change in ambient tem-

perature, electrostatic discharge, formation of electrical charges (triboelectric effect), vibration of a connecting cable, electromagnetic interference (EMI), and so forth. Even if a manufacturer does not specify such effects, an application engineer should simulate them during the prototype phase of the design process. If, indeed, the environmental factors degrade the sensor's performance, additional corrective measures may be required (see Chapter 4) (e.g., placing the sensor in a protective box, using electrical shielding, using a thermal insulation or a thermostat).

Temperature factors are very important for sensor performance; they must be known and taken into account. The operating temperature range is the span of ambient temperatures given by their upper and lower extremes (e.g., $-20°C$ to $+100°C$) within which the sensor maintains its specified accuracy. Many sensors change with temperature and their transfer functions may shift significantly. Special compensating elements are often incorporated either directly into the sensor or into signal conditioning circuits, to compensate for temperature errors. The simplest way of specifying tolerances of thermal effects is provided by the error-band concept, which is simply the error band that is applicable over the operating temperature band. A temperature band may be divided into sections, whereas the error band is separately specified for each section. For example, a sensor may be specified to have an accuracy of $\pm 1\%$ in the range from $0°C$ to $50°C$, $\pm 2\%$ from $-20°C$ to $0°C$ and from $+50°C$ to $100°C$, and $\pm 3\%$ beyond these ranges within operating limits specified from $-40°C$ to $+150°C$.

Temperatures will also affect dynamic characteristics, particularly when they employ viscous damping. A relatively fast temperature change may cause the sensor to generate a spurious output signal. For instance, a dual pyroelectric sensor in a motion detector is insensitive to slowly varying ambient temperature. However, when the temperature changes quickly, the sensor will generate an electric current that may be recognized by a processing circuit as a valid response to a stimulus, thus causing a false-positive detection.

A *self-heating error* may be specified when an excitation signal is absorbed by a sensor and changes its temperature by such a degree that it may affect its accuracy. For instance, a thermistor temperature sensor requires passage of electric current, causing heat dissipation within the sensor's body. Depending on its coupling with the environment, the sensors' temperature may increase due to a self-heating effect. This will result in errors in temperature measurement because the thermistor now acts as an additional spurious source of thermal energy. The coupling depends on the media in which the sensor operates—a dry contact, liquid, air, and so forth. A worst coupling may be through still air. For thermistors, manufacturers often specify self-heating errors in air, stirred liquid, or other media.

A sensor's temperature increase above its surroundings may be found from the following formula:

$$\Delta T^{\circ} = \frac{V^2}{(\xi v c + \alpha) R}, \tag{2.25}$$

where ξ is the sensor's mass density, c is specific heat, v is the volume of the sensor, α is the coefficient of thermal coupling between the sensor and the outside (thermal conductivity), R is the electrical resistance, and V is the effective voltage across the resistance. If a self-heating results in an error, Eq. (2.25) may be used as a design

guide. For instance, to increase α, a thermistor detector should be well coupled to the object by increasing the contact area, applying thermally conductive grease or using thermally conductive adhesives. Also, high-resistance sensors and low measurement voltages are preferable.

2.18 Reliability

Reliability is the ability of a sensor to perform a required function under stated conditions for a stated period. It is expressed in statistical terms as a probability that the device will function without failure over a specified time or a number of uses. It should be noted that reliability is not a characteristic of drift or noise stability. It specifies a *failure*, either temporary or permanent, exceeding the limits of a sensor's performance under normal operating conditions.

Reliability is an important requirement; however, it is rarely specified by the sensor manufacturers. Probably, the reason for that is the absence of a commonly accepted measure for the term. In the United States, for many electronic devices, the procedure for predicting in-service reliability is the MTBF (mean time between failure) calculation described in MIL-HDBK-217 standard. Its basic approach is to arrive at a MTBF rate for a device by calculating the individual failure rates of the individual components used and by factoring in the kind of operation the device will see: its temperature, stress, environment, and screening level (measure of quality). Unfortunately, the MTBF reflects reliability only indirectly and it is often hardly applicable to everyday use of the device. The qualification tests on sensors are performed on combinations of the worst possible conditions. One approach (suggested by MIL-STD-883) is 1000 h, loaded at maximum temperature. This test does not qualify for such important impacts as fast temperature changes. The most appropriate method of testing would be accelerated life qualification. It is a procedure that emulates the sensor's operation, providing real-world stresses, but compressing years into weeks. Three goals are behind the test: to establish MTBF; to identify first failure points that can then be strengthened by design changes; and to identify the overall system practical lifetime.

One possible way to compress time is to use the same profile as the actual operating cycle, including maximum loading and power-on, power-off cycles, but expanded environmental highest and lowest ranges (temperature, humidity, and pressure). The highest and lowest limits should be substantially broader than normal operating conditions. Performance characteristics may be outside specifications, but must return to those when the device is brought back to the specified operating range. For example, if a sensor is specified to operate up to 50°C at the highest relative humidity (RH) of 85% at a maximum supply voltage of +15 V, it may be cycled up to 100°C at 99% RH and at +18 V power supply. To estimate number of test cycles (n), the following empirical formula [developed by Sandstrand Aerospace, (Rockford, IL) and Interpoint Corp. (Redmond, WA)] [1] may be useful:

$$n = N\left(\frac{\Delta T_{\max}}{\Delta T_{\text{test}}}\right)^{2.5}, \tag{2.26}$$

where N is the estimated number of cycles per lifetime, $\Delta T_{\max}$ is the maximum specified temperature fluctuation, and ΔT_{test} maximum cycled temperature fluctuation during the test. For instance, if the normal temperature is 25°C, the maximum specified temperature is 50°C, cycling was up to 100°C, and over the lifetime (say, 10 years), the sensor was estimated to be subjected to 20,000 cycles, then the number of test cycles is calculated as

$$n = 20{,}000 \left(\frac{50 - 25}{100 - 25} \right)^{2.5} = 1283.$$

As a result, the accelerated life test requires about 1300 cycles instead of 20,000. It should be noted, however, that the 2.5 factor was derived from a solder fatigue multiple, because that element is heavily influenced by cycling. Some sensors have no solder connections at all, and some might have even more sensitivity to cycling substances other than solder, (e.g, electrically conductive epoxy). Then, the factor should be selected to be somewhat smaller. As a result of the accelerated life test, the reliability may be expressed as a probability of failure. For instance, if 2 out of 100 sensors (with an estimated lifetime of 10 years) failed the accelerated life test, the reliability is specified as 98% over 10 years.

A sensor, depending on its application, may be subjected to some other environmental effects which potentially can alter its performance or uncover hidden defects. Among such additional tests are:

- High temperature/high humidity while being fully electrically powered. For instance, a sensor may be subjected to its maximum allowable temperature at 85–90% RH and kept under these conditions for 500 h. This test is very useful for detecting contaminations and evaluating packaging integrity. The life of sensors, operating at normal room temperatures, is often accelerated at 85°C and 85% RH, which is sometimes called an "85–85 test."
- Mechanical shocks and vibrations may be used to simulate adverse environmental conditions, especially in the evaluation wire bonds, adhesion of epoxy, and so forth. A sensor may be dropped to generate high-level accelerations (up to 3000g of force). The drops should be made on different axes. Harmonic vibrations should be applied to the sensor over the range which includes its natural frequency. In the United States military standard 750, methods 2016 and 2056 are often used for mechanical tests.
- Extreme storage conditions may be simulated, for instance at +100 and −40°C while maintaining a sensor for at least 1000 h under these conditions. This test simulates storage and shipping conditions and usually is performed on nonoperating devices. The upper and lower temperature limits must be consistent with the sensor's physical nature. For example, TGS pyroelectric sensors manufactured in the past by Philips are characterized by a Curie temperature of +60°C. Approaching and surpassing this temperature results in a permanent destruction of sensitivity. Hence, the temperature of such sensors should never exceed +50°C, which must be clearly specified and marked on its packaging material.

- Thermal shock or temperature cycling (TC) is subjecting a sensor to alternate extreme conditions. For example, it may be dwelled for 30 min at −40°C, then quickly moved to +100°C for 30 min, and then back to cold. The method must specify the total number of cycling, like 100 or 1000. This test helps to uncover die bond, wire bond, epoxy connections, and packaging integrity.
- To simulate sea conditions, sensors may be subjected to a salt spray atmosphere for a specified time, (e.g., 24 h). This helps to uncover its resistance to corrosion and structural defects.

2.19 Application Characteristics

Design, *weight*, and overall *dimensions* are geared to specific areas of applications. *Price* may be a secondary issue when the sensor's reliability and accuracy are of paramount importance. If a sensor is intended for life-support equipment, weapons or spacecraft, a high price tag may be well justified to assure high accuracy and reliability. On the other hand, for a very broad range of consumer applications, the price of a sensor often becomes a cornerstone of a design.

2.20 Uncertainty

Nothing is perfect in this world, at least in the sense that we perceive it. All materials are not exactly as we think they are. Our knowledge of even the purest of the materials is always approximate; machines are not perfect and never produce perfectly identical parts according to drawings. All components experience drifts related to the environment and their aging; external interferences may enter the system and alter its performance and modify the output signal. Workers are not consistent and the human factor is nearly always present. Manufacturers fight an everlasting battle for the uniformity and consistency of the processes, yet the reality is that every part produced is never ideal and carries an uncertainty of its properties. Any measurement system consists of many components, including sensors. Thus, no matter how accurate the measurement is, it is only an approximation or estimate of the true value of the specific quantity subject to measurement, (i.e., the stimulus or measurand). The result of a measurement should be considered complete only when accompanied by a quantitative statement of its uncertainty. We simply never can be 100% sure of the measured value.

When taking individual measurements (samples) under noisy conditions we expect that the stimulus s is represented by the sensor as having a somewhat different value s', so that the error in measurement is expressed as

$$\delta = s' - s, \tag{2.27}$$

The difference between the *error* specified by Eq. (2.27) and *uncertainty* should always be clearly understood. An error can be compensated to a certain degree by correcting its systematic component. The result of such a correction can unknowably be very close to the unknown true value of the stimulus and, thus, it will have a very

small error. Yet, in spite of a small error, the uncertainty of measurement may be very large so we cannot really trust that the error is indeed that small. In other words, an error is what we unknowably *get* when we measure, whereas uncertainty is what we *think* how large that error might be.

The International Committee for Weight and Measures *(CIPM)* considers that uncertainty consists of many factors that can be grouped into two classes or types [2,3]:

A: Those evaluated by statistical methods
B: Those evaluated by other means.

This division is not clear-cut and the borderline between Types A and B is somewhat illusive. Generally, Type A components of uncertainty arise from random effects, whereas the Type B components arise from systematic effects.

Type A uncertainty is generally specified by a standard deviation s_i, equal to the positive square root of the statistically estimated variance s_i^2 and the associated number of degrees of freedom v_i. For such a component, the *standard* uncertainty is $u_i = s_i$. Standard uncertainty represents each component of uncertainty that contributes to the uncertainty of the measurement result.

The evaluation of a Type A standard uncertainty may be based on any valid statistical method for treating data. Examples are calculating the standard deviation of the mean of a series of independent observations, using the method of least squares to fit a curve to data in order to estimate the parameters of the curve and their standard deviations. If the measurement situation is especially complicated, one should consider obtaining the guidance of a statistician.

The evaluation of a Type B standard uncertainty is usually based on scientific judgment using all of the relevant information available, which may include the following:

- Previous measurement data
- Experience with or general knowledge of the behavior and property of relevant sensors, materials, and instruments
- Manufacturer's specifications
- Data obtained during calibration and other reports
- Uncertainties assigned to reference data taken from handbooks and manuals

For detailed guidance of assessing and specifying standard uncertainties one should consult specialized texts (e.g., Ref. [4]).

When both Type A and Type B uncertainties are evaluated, they should be combined to represent the *combined standard uncertainty*. This can be done by using a conventional method for combining standard deviations. This method is often called the *law of propagation of uncertainty* and in common parlance is known as "root-sum-of-squares" (square root of the sum-of-the-squares) or RSS method of combining uncertainty components estimated as standard deviations:

$$u_c = \sqrt{u_1^2 + u_2^2 + \cdots + u_i^2 + \cdots + u_n^2}, \tag{2.28}$$

where n is the number of standard uncertainties in the uncertainty budget.

Table 2.2. Uncertainty Budget for Thermistor Thermometer

Source of Uncertainty	Standard uncertainty (°C)	Type
Calibration of sensor	0.03	B
Measured errors		
Repeated observations	0.02	A
Sensor noise	0.01	A
Amplifier noise	0.005	A
Sensor aging	0.025	B
Thermal loss through connecting wires	0.015	A
Dynamic error due to sensor's inertia	0.005	B
Temperature instability of object of measurement	0.04	A
Transmitted noise	0.01	A
Misfit of transfer function	0.02	B
Ambient drifts		
Voltage reference	0.01	A
Bridge resistors	0.01	A
Dielectric absorption in A/D capacitor	0.005	B
Digital resolution	0.01	A
Combined standard uncertainty	**0.068**	

Table 2.2 shows an example of an uncertainty budget for an electronic thermometer with a thermistor sensor which measures the temperature of a water bath. While compiling such a table, one must be very careful not to miss any standard uncertainty, not only in a sensor but also in the interface instrument, experimental setup, and the object of measurement. This must be done for various environmental conditions, which may include temperature, humidity, atmospheric pressure, power supply variations, transmitted noise, aging, and many other factors.

No matter how accurately any individual measurement is made, (i.e., how close the measured temperature is to the true temperature of an object), one never can be sure that it is indeed accurate. The combined standard uncertainty of 0.068°C does not mean that the error of measurement is no greater than 0.068°C. That value is just a standard deviation, and if an observer has enough patience, he may find that individual errors may be much larger. The word "uncertainty" by its very nature implies that the uncertainty of the result of a measurement is an estimate and generally does not have well-defined limits.

References

1. Better reliability via system tests. *Electron. Eng. Times* 40–41, Aug. 19, 1991.
2. CIPM, *BIPM Proc.-Verb. Com. Int. Poids et Mesures* 49, pp. 8–9, No. 26, 1981 (in French).

3. *ISO Guide to the Expression of Uncertainty in Measurements*. International Organization for Standardization, Geneva, 1993.
4. Taylor, B. N. and Kuyatt, C. E. *Guidelines for Evaluation and Expressing the Uncertainty of NIST Measurement Results*. NIST Technical Note 1297, Gaithersburg, 1994.

3

Physical Principles of Sensing

"The way we have to describe Nature
is generally incomprehensible to us."

—Richard P. Feynman,
"QED. The Strange Theory of Light and Matter"

"It should be possible to explain
the laws of physics to a barmaid."

—Albert Einstein

Because a sensor is a converter of generally nonelectrical effects into electrical signals, one and often several transformation steps are required before the electric output signal can be generated. These steps involve changes of the types of energy, where the final step must produce an electrical signal of a desirable format. As was mentioned in Chapter 1, generally there are two types of sensor: *direct* and *complex*. A direct sensor is the one that can directly convert a nonelectrical stimulus into an electric signal. Many stimuli cannot be directly converted into electricity, thus multiple conversion steps would be required. If, for instance, one wants to detect the displacement of an opaque object, a fiber-optic sensor can be employed. A pilot (excitation) signal is generated by a light-emitting diode (LED), transmitted via an optical fiber to the object and reflected from its surface. The reflected photon flux enters the receiving optical fiber and propagates toward a photodiode, where it produces an electric current representing the distance from the fiber-optic end to the object. We see that such a sensor involves the transformation of electrical current into photons, the propagation of photons through some refractive media, reflection, and conversion back into electric current. Therefore, such a sensing process includes two energy-conversion steps and a manipulation of the optical signal as well.

There are several physical effects which result in the direct generation of electrical signals in response to nonelectrical influences and thus can be used in direct sensors. Examples are thermoelectric (Seebeck) effect, piezoelectricity, and photoeffect.

This chapter examines various physical effects that can be used for a *direct* conversion of stimuli into electric signals. Because all such effects are based on fundamental principles of physics, we briefly review these principles from the standpoint of sensor technologies.

3.1 Electric Charges, Fields, and Potentials

There is a well-known phenomenon to those who live in dry climates—the possibility of the generation of sparks by friction involved in walking across the carpet. This is a result of the so-called *triboelectric effect*,[1] which is a process of an electric charge separation due to object movements, friction of clothing fibers, air turbulence, atmosphere electricity, and so forth. There are two kinds of charge. Like charges repel each other and the unlike charges attract each other. Benjamin Franklin (1706–1790), among his other remarkable achievements, was the first American physicist. He named one charge *negative* and the other *positive*. These names have remained to this day. He conducted an elegant experiment with a kite flying in a thunderstorm to prove that the atmospheric electricity is of the same kind as produced by friction. In doing the experiment, Franklin was extremely lucky, as several Europeans who were trying to repeat his test were severely injured by the lightning and one was killed.

A triboelectric effect is a result of a mechanical charge redistribution. For instance, rubbing a glass rod with silk strips electrons from the surface of the rod, thus leaving an abundance of positive charges (i.e., giving the rod a positive charge). It should be noted that the electric charge is conserved: It is neither created nor destroyed. Electric charges can be only moved from one place to another. Giving negative charge means taking electrons from one object and placing them onto another (charging it negatively). The object which loses some amount of electrons is said gets a positive charge.

A triboelectric effect influences an extremely small number of electrons as compared with the total electronic charge in an object. The actual amount of charges in any object is very large. To illustrate this, let us consider the total number of electrons in a U.S. copper penny[2] [1]. The coin weighs 3.1 g; therefore, it can be shown that the total number of atoms in it is about 2.9×10^{22}. A copper atom has a positive nuclear charge of 4.6×10^{-18} C and the same electronic charge of the opposite polarity. A combined charge of all electrons in a penny is $q = (4.6 \times 10^{-18}\text{C/atom})(2.9 \times 10^{22}\text{atoms}) = 1.3 \times 10^{5}$ C, a very large charge indeed. This electronic charge from a single copper penny may generate a sufficient current of 0.91 A to operate a 100-W light bulb for 40 h.

With respect to electric charges, there are three kinds of material: conductors, isolators, and semiconductors. In conductors, electric charges (electrons) are free to move through the material, whereas in isolators, they are not. Although there is no

[1] The prefix *tribo* means "pertinent to friction."

[2] Currently, the U.S. pennies are just copper-plated zinc alloy, but before 1982 they were made of copper.

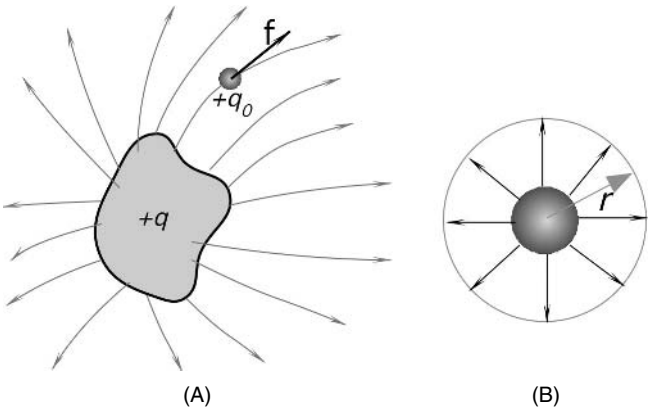


Fig. 3.1. (A) Positive test charge in the vicinity of a charged object and (B) the electric field of a spherical object.

perfect isolator, the isolating ability of fused quartz is about 10^{25} times as great as that of copper, so that for practical purposes, many materials are considered perfect isolators. The semiconductors are intermediate between conductors and isolators in their ability to conduct electricity. Among the elements, silicon and germanium are well-known examples. In semiconductors, the electrical conductivity may be greatly increased by adding small amounts of other elements; traces of arsenic or boron are often added to silicon for this purpose.

Figure 3.1A shows an object which carries a positive electric charge q. If a small *positive* electric test charge q_0 is positioned in the vicinity of a charged object, it will be subjected to a repelling electric force. If we place a negative charge on the object, it will attract the test charge. In vector form, the repelling (or attracting) force is shown as $\mathbf{f}$. The boldface indicates a vector notation. A fact that the test charge is subjected to force without a physical contact between charges means that the volume of space occupied by the test charge may be characterized by a so-called *electric field*.

The electric field in each point is defined through the force as

$$\mathbf{E} = \frac{\mathbf{f}}{q_0}. \tag{3.1}$$

Here, $\mathbf{E}$ is vector in the same direction as $\mathbf{f}$ because q_0 is scalar. Formula (3.1) expresses an electric field as a force divided by a property of a test charge. The test charge must be very small not to disturb the electric field. Ideally, it should be infinitely small; however, because the charge is quantized, we cannot contemplate a free test charge whose magnitude is smaller than the electronic charge: $e = 1.602 \times 10^{-19}$ C.

The field is indicated in Fig. 3.1A by the *field lines* which in every point of space are tangent to the vector of force. By definition, the field lines start on the positive plate and end on the negative. The density of field lines indicates the magnitude of the electric field $\mathbf{E}$ in any particular volume of space.

For a physicist, any field is a physical quantity that can be specified simultaneously for all points within a given region of interest. Examples are pressure field, temperature fields, electric fields, and magnetic fields. A field variable may be a scalar (e.g., temperature field) or a vector (e.g., a gravitational field around the Earth). The field variable may or may not change with time. A vector field may be characterized by a distribution of vectors which form the so-called flux (Φ). Flux is a convenient description of many fields, such as electric, magnetic, thermal, and so forth. The word "flux" is derived from the Latin word *fluere* (to flow). A familiar analogy of flux is a stationary, uniform field of fluid flow (water) characterized by a constant flow vector **v**, the constant velocity of the fluid at any given point. In case of an electric field, nothing flows in a formal sense. If we replace **v** by **E** (vector representing the electric field), the field lines form flux. If we imagine a hypothetical closed surface (Gaussian surface) ***S***, a connection between the charge q and flux can be established as

$$\varepsilon_0 \Phi_{\mathrm{E}} = q, \tag{3.2}$$

where $\varepsilon_0 = 8.8542 \times 10^{-12}$ C^2/N m^2 is the permitivity constant, or by integrating flux over the surface,

$$\varepsilon_0 \oint \mathbf{E}\, ds = q, \tag{3.3}$$

where the integral is equal to Φ_E. In the above equations, known as Gauss' law, the charge q is the net charge surrounded by the Gaussian surface. If a surface encloses equal and opposite charges, the net flux Φ_E is zero. The charge outside the surface makes no contribution to the value of q, nor does the exact location of the inside charges affect this value. Gauss' law can be used to make an important prediction, namely *an exact charge on an insulated conductor is in equilibrium, entirely on its outer surface*. This hypothesis was shown to be true even before either Gauss' law or Coulomb's law was advanced. Coulomb's law itself can be derived from Gauss' law. It states that the force acting on a test charge is inversely proportional to a squared distance from the charge:

$$f = \frac{1}{4\pi\varepsilon_0} \frac{q q_0}{r^2}. \tag{3.4}$$

Another result of Gauss' law is that the electric field outside any spherically symmetrical distribution of charge (Fig. 3.1B) is directed radially and has magnitude (note that magnitude is not a vector)

$$E = \frac{1}{4\pi\varepsilon_0} \frac{q}{r^2}, \tag{3.5}$$

where r is the distance from the sphere's center.

Similarly, the electric field inside a uniform sphere of charge q is directed radially and has magnitude

$$E = \frac{1}{4\pi\varepsilon_0} \frac{qr}{R^3}, \tag{3.6}$$

where R is the sphere's radius and r is the distance from the sphere's center. It should be noted that the electric field in the center of the sphere ($r = 0$) is equal to zero.

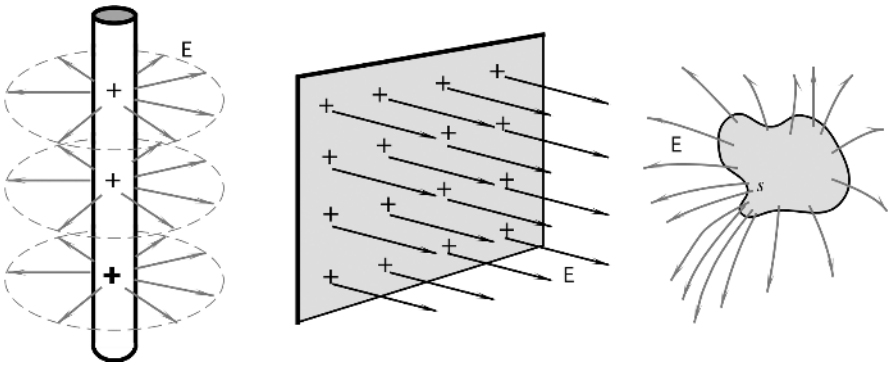


Fig. 3.2. Electric field around an infinite line (A) and near an infinite sheet (B). A pointed conductor concentrates an electric field (C).

If the electric charge is distributed along an infinite (or, for the practical purposes, long) line (Fig. 3.2A), the electric field is directed perpendicularly to the line and has the magnitude

$$E = \frac{\lambda}{2\pi\varepsilon_0 r}, \tag{3.7}$$

where r is the distance from the line and λ is the linear charge density (charge per unit length). The electric field due to an infinite sheet of charge (Fig. 3.2B) is perpendicular to the plane of the sheet and has magnitude

$$E = \frac{\sigma}{2\varepsilon_0}, \tag{3.8}$$

where σ is the surface charge density (charge per unit area). However, for an isolated conductive object, the electric field is two times stronger:

$$E = \frac{\sigma}{\varepsilon_0}. \tag{3.9}$$

The apparent difference between electric fields of Eqs. (3.8) and (3.9) is a result of different geometries: The former is an infinite sheet and the latter is an object of an arbitrary shape. A very important consequence of Gauss' law is that electric charges are distributed only on the outside surface. This is a result of repelling forces between charges of the same sign: All charges try to move as far as possible from one another. The only way to do this is to move to the foremost distant place in the material, which is the outer surface. Of all places on the outer surface, the most preferable places are the areas with the highest curvatures. This is why pointed conductors are the best concentrators of the electric field (Fig. 3.2C). A very useful scientific and engineering tool is a Faraday cage: a room entirely covered by either grounded conductive sheets or a metal net. No matter how strong the external electric field, it will be essentially zero inside the cage. This makes cars and metal ships the best

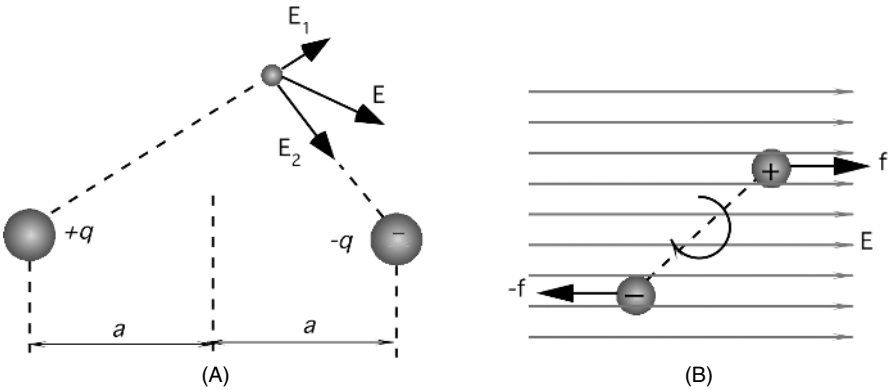


Fig. 3.3. Electric dipole (A); an electric dipole in an electric field is subjected to a rotating force (B).

protectors during thunderstorms, because they act as virtual Faraday cages. It should be remembered, however, that the Faraday cage, although being a perfect shield against electric fields, is of little use to protect against magnetic fields, unless it is made of a thick ferromagnetic material.

An *electric dipole* is a combination of two opposite charges placed at a distance $2a$ apart (Fig. 3.3A). Each charge will act on a test charge with force which defines electric fields $\mathbf{E_1}$ and $\mathbf{E_2}$ produced by individual charges. A combined electric field of a dipole, $\mathbf{E}$, is a vector sum of two fields. The magnitude of the field is

$$E = \frac{1}{4\pi\varepsilon_0}\frac{qa}{r^3}, \tag{3.10}$$

where r is the distance from the center of the dipole. The essential properties of the charge distribution are the magnitude of the charge q and the separation $2a$. In formula (3.10), the charge and distance are entered only as a product. This means that if we measure E at various distances from the electric dipole (assuming that the distance is much longer than a), we can never deduce q and $2a$ separately, but only the product $2qa$. For instance, if q is doubled and a is cut in half, the electric field will not change. The product $2qa$ is called the electric dipole moment p. Thus, Eq. (3.10) can be rewritten as

$$E = \frac{1}{4\pi\varepsilon_0}\frac{p}{r^3}. \tag{3.11}$$

The spatial position of a dipole may be specified by its moment in vector form: $\mathbf{p}$. Not all materials have a dipole moment: Gases such as methane, acetylene, ethylene, carbon dioxide, and many others have no dipole moment. On the other hand, carbon monoxide has a weak dipole moment (0.37×10^{-30}C m) and water has a strong dipole moment (6.17×10^{-30}C m).

Dipoles are found in crystalline materials and form a foundation for such sensors as piezoelectric and pyroelectric detectors. When a dipole is placed in an electric field,

it becomes subjected to a rotation force (Fig. 3.3B). Usually, a dipole is a part of a crystal which defines its initial orientation. An electric field, if strong enough, will align the dipole along its lines. Torque, which acts on a dipole in a vector form, is

$$\tau = \mathbf{pE}. \tag{3.12}$$

Work must be done by an external agent to change the orientation of an electric dipole in an external electric field. This work is stored as potential energy U in the system consisting of the dipole and the arrangement used to set up the external field. In a vector form this potential energy is

$$U = -\mathbf{pE}. \tag{3.13}$$

A process of dipole orientation is called *poling*. The aligning electric field must be strong enough to overcome a retaining force in the crystalline stricture of the material. To ease this process, the material during the poling is heated to increase the mobility of its molecular structure. The poling is used in fabrication of piezoelectric and pyroelectric crystals.

The electric field around the charged object can be described not only by the vector **E**, but by a scalar quantity, the *electric potential V* as well. Both quantities are intimately related and usually it is a matter of convenience which one to use in practice. A potential is rarely used as a description of an electric field in a specific point of space. A potential difference (voltage) between two points is the most common quantity in electrical engineering practice. To find the voltage between two arbitrary points, we may use the same technique as above—a small positive test charge q_0. If the electric charge is positioned in point A, it stays in equilibrium, being under the influence of force $q_0\mathbf{E}$. Theoretically, it may remain there infinitely long. Now, if we try to move it to another point B, we have to work against the electric field. Work (W_{AB}) which is done against the field (that is why it has negative sign) to move the charge from A to B defines the voltage between these two points:

$$V_B - V_A = -\frac{W_{AB}}{q_0}. \tag{3.14}$$

Correspondingly, the electrical potential at point B is smaller than at point A. The SI unit for voltage is 1 volt = 1 joule/coulomb. For convenience, point A is chosen to be very far away from all charges (theoretically at an infinite distance) and the electric potential at that point is considered to be zero. This allows us to define the electric potential at any other point as

$$V = -\frac{W}{q_0}. \tag{3.15}$$

This equation tells us that the potential near the positive charge is positive, because moving the positive test charge from infinity to the point in a field, must be made against a repelling force. This will cancel the negative sign in formula (3.15). It should be noted that the potential difference between two points is independent of the path along which the test charge is moving. It is strictly a description of the electric field

difference between the two points. If we travel through the electric field along a straight line and measure V as we go, the rate of change of V with distance l that we observe is the components of **E** in that direction

$$E_l = -\frac{dV}{dl}. \tag{3.16}$$

The minus sign tells us that **E** points in the direction of decreasing V. Therefore, the appropriate units for electric field is volts/meter (V/m).

3.2 Capacitance

Let us take two isolated conductive objects of arbitrary shape (plates) and connect them to the opposite poles of a battery (Fig. 3.4A). The plates will receive equal amounts of opposite charges; that is, a negatively charged plate will receive additional electrons while there will be a deficiency of electrons in the positively charged plate. Now, let us disconnect the battery. If the plates are totally isolated and exist in a vacuum, they will remain charged theoretically infinitely long. A combination of plates which can hold an electric charge is called a *capacitor*. If a small *positive* electric test charge, q_0, is positioned between the charged objects, it will be subjected to an electric force from the positive plate to the negative. The positive plate will repel the test charge and the negative plate will attract it, resulting in a combined push-pull force. Depending on the position of the test charge between the oppositely charged objects, the force will have a specific magnitude and direction, which is characterized by vector **f**.

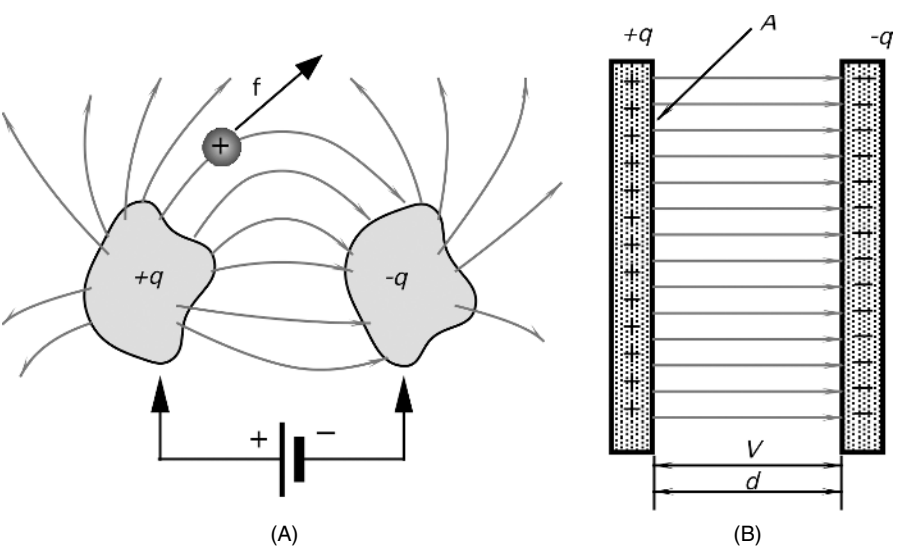


Fig. 3.4. Electric charge and voltage define the capacitance between two objects (A); a parallel-plate capacitor (B).

The capacitor may be characterized by q, the magnitude of the charge on either conductor (shown in Fig. 3.4A), and by V, the positive potential difference between the conductors. It should be noted that q is not a net charge on the capacitor, which is zero. Further, V is not the potential of either plate, but the potential difference between them. The ratio of charge to voltage is constant for each capacitor:

$$\frac{q}{V} = C. \tag{3.17}$$

This fixed ratio, C, is called the *capacitance* of the capacitor. Its value depends on the shapes and relative position of the plates. C also depends on the medium in which the plates are immersed. Note that C is always positive because we use the same sign for both q and V. The SI unit for capacitance is 1 farad = 1 coulomb/volt, which is represented by the abbreviation F. A farad is a very large capacitance; hence, in practice submultiples of the farad are generally used:

1 picofarad (pF) $= 10^{-12}$ F
1 nanofarad (nF) $= 10^{-9}$ F
1 microfarad (μF) $= 10^{-6}$ F

When connected into an electronic circuit, capacitance may be represented as a "complex resistance":

$$\frac{V}{i} = -\frac{1}{j\omega C}, \tag{3.18}$$

where $j = \sqrt{-1}$ and i is the sinusoidal current having a frequency of ω, meaning that the complex resistance of a capacitor drops at higher frequencies. This is called Ohm's law for the capacitor. The minus sign and complex argument indicate that the voltage across the capacitor lags 90° behind the current.

Capacitance is a very useful physical phenomenon in a sensor designer's toolbox. It can be successfully applied to measure distance, area, volume, pressure, force, and so forth. The following background establishes fundamental properties of the capacitor and gives some useful equations. Figure 3.4B shows a parallel-plate capacitor in which the conductors take the form of two plane parallel plates of area A separated by a distance d. If d is much smaller than the plate dimensions, the electric field between the plates will be uniform, which means that the field lines (lines of force **f**) will be parallel and evenly spaced. The laws of electromagnetism requires that there be some "fringing" of the lines at the edges of the plates, but for small enough d, we can neglect it for our present purpose.

3.2.1 Capacitor

To calculate the capacitance, we must relate V, the potential difference between the plates, to q, the capacitor charge (3.17):

$$C = \frac{q}{V}. \tag{3.19}$$

Alternatively, the capacitance of a flat capacitor can be found from

$$C = \frac{\varepsilon_0 A}{d}. \tag{3.20}$$

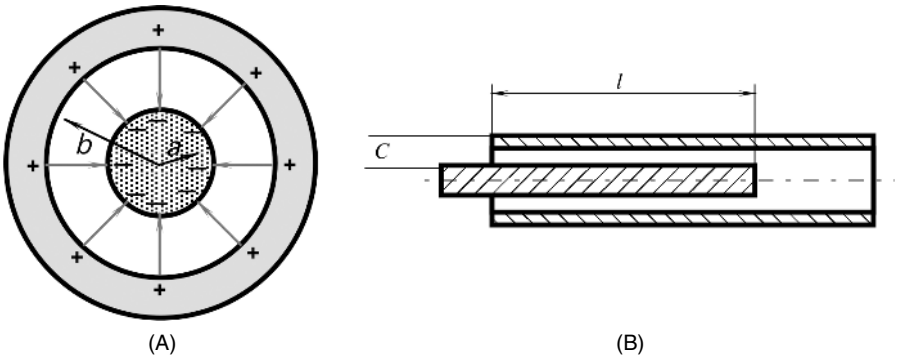


Fig. 3.5. Cylindrical capacitor (A); capacitive displacement sensor (B).

Formula (3.20) is important for the capacitive sensor's design. It establishes a relationship between the plate area and the distance between the plates. Varying either of them will change the capacitor's value, which can be measured quite accurately by an appropriate circuit. It should be noted that Eqs. (3.19) and (3.20) hold only for capacitors of the parallel type. A change in geometry will require modified formulas. The ratio A/d may be called a geometry factor for a parallel-plate capacitor.

A cylindrical capacitor, shown in Fig. 3.5A, consists of two coaxial cylinders of radii a and b and length l. For the case when $l \gg b$, we can ignore fringing effects and calculate capacitance from the following formula:

$$C = \frac{2\pi\varepsilon_0 l}{\ln(b/a)}. \tag{3.21}$$

In this formula, l is the length of the overlapping conductors (Fig. 3.5B) and $2\pi l[\ln(b/a)]^{-1}$ is called a geometry factor for a coaxial capacitor. A useful displacement sensor can be built with such a capacitor if the inner conductor can be moved in and out of the outer conductor. According to Eq. (3.21), the capacitance of such a sensor is in a linear relationship with the displacement, l.

3.2.2 Dielectric Constant

Equation (3.20) holds for a parallel-plate capacitor with its plates in vacuum (or air, for most practical purposes). In 1837, Michael Faraday first investigated the effect of completely filling the space between the plates with a dielectric. He had found that the effect of the filling is to increase the capacitance of the device by a factor of κ, which is known as the dielectric constant of the material.

The increase in capacitance due to the dielectric presence is a result of molecular polarization. In some dielectrics (e.g., in water), molecules have a permanent dipole moment, whereas in other dielectrics, molecules become polarized only when an external electric field is applied. Such a polarization is called induced. Both cases, either permanent electric dipoles or those acquired by induction, tend to align molecules

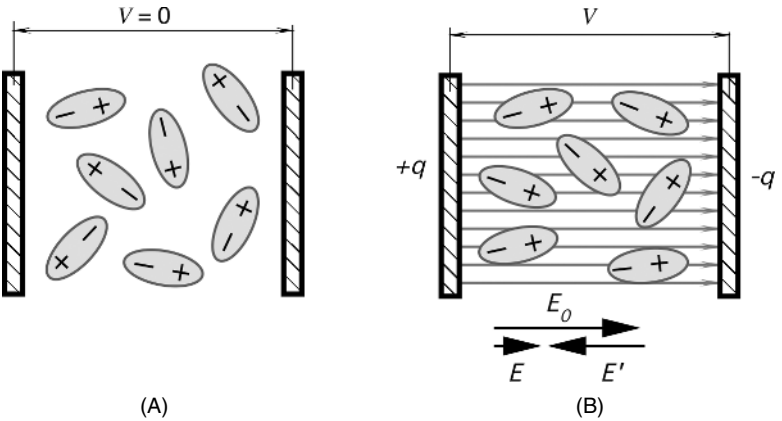


Fig. 3.6. Polarization of dielectric: (A) dipoles randomly oriented without an external electric field; (B) dipoles aligned with an electric field.

with an external electric field. This process is called dielectric polarization. It is illustrated in Fig. 3.6A which shows permanent dipoles before an external electric field is applied to the capacitor, and in Fig. 3.6B, which shows permanent dipoles after an external electric field is applied to the capacitor. In the former case, there is no voltage between the capacitor plates, and all dipoles are randomly oriented. After the capacitor is charged, the dipoles will align with the electric field lines; however, thermal agitation will prevent a complete alignment. Each dipole forms its own electric field which is predominantly oppositely directed with the external electric field, $\mathbf{E_0}$. Due to a combined effect of a large number of dipoles ($\mathbf{E'}$), the electric field in the capacitor becomes weaker ($\mathbf{E} = \mathbf{E_0} + \mathbf{E'}$) when the field, $\mathbf{E_0}$, would be in the capacitor without the dielectric.

Reduced electric field leads to a smaller voltage across the capacitor: $V = V_0/\kappa$. Substituting it into formula (3.19), we get an expression for the capacitor with a dielectric:

$$C = \kappa \frac{q}{V_0} = \kappa C_0. \tag{3.22}$$

For the parallel-plate capacitor, we thus have

$$C = \frac{\kappa \varepsilon_0 A}{d}. \tag{3.23}$$

In a more general form, the capacitance between two objects may be expressed through a geometry factor, G:

$$C = \varepsilon_0 \kappa G, \tag{3.24}$$

G depends on the shape of the objects (plates) and their separation. Table A.5 of the Appendix gives the dielectric constants, κ, for various materials.

Dielectric constants must be specified for test frequency and temperature. Some dielectrics have a very uniform dielectric constant over a broad frequency range (e.g.,

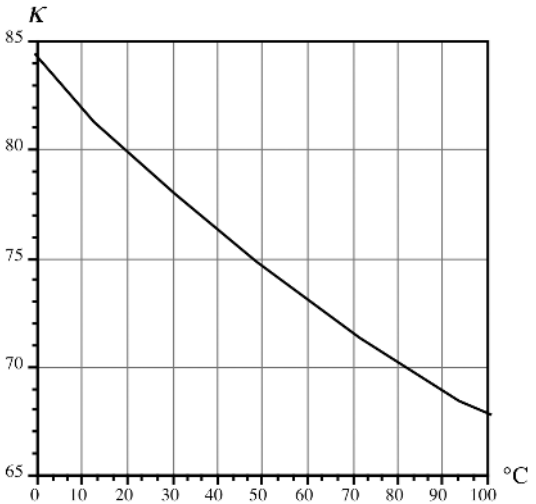


Fig. 3.7. Dielectric constant of water as a function of temperature.

polyethylene), whereas others display a strong negative frequency dependence; that is, a dielectric constant decreases with frequency. The temperature dependence is also negative. Figure 3.7 illustrates κ for water as a function of temperature.

In a "good" capacitor, a dielectric constant κ and geometry must be stable—ideally, they should not vary with temperature, humidity, pressure, or any other environmental factors. "Good" capacitors are essential components of electronic circuits. However, if you want to design a capacitive sensor, you need to make a "bad" capacitor, whose value varies with temperature, humidity, pressure, or whatever you need to sense. By allowing a capacitor's parameter to vary *selectively* with a specific stimulus, one can build a useful sensor.

Let us consider a capacitive water-level sensor (Fig. 3.8A). The sensor is fabricated in a form of a coaxial capacitor where the surface of each conductor is coated with a thin isolating layer to prevent an electric short circuit through water (the isolator is a dielectric which we disregard in the following analysis because it does not change in the process of measurement). The sensor is immersed in a water tank. When the level increases, water fills more and more space between the sensor's coaxial conductors, thus changing the sensor's capacitance. The total capacitance of the coaxial sensor is

$$C_h = C_1 + C_2 = \varepsilon_0 G_1 + \varepsilon_0 \kappa G_2, \tag{3.25}$$

where C_1 is the capacitance of the water-free portion of the sensor and C_2 is the capacitance of the water-filled portion. The corresponding geometry factors are designated G_1 and G_2. From formulas (3.21) and (3.25), the total sensor capacitance can be found as

$$C_h = \frac{2\pi\varepsilon_0}{\ln(b/a)}[H - h(1-\kappa)], \tag{3.26}$$

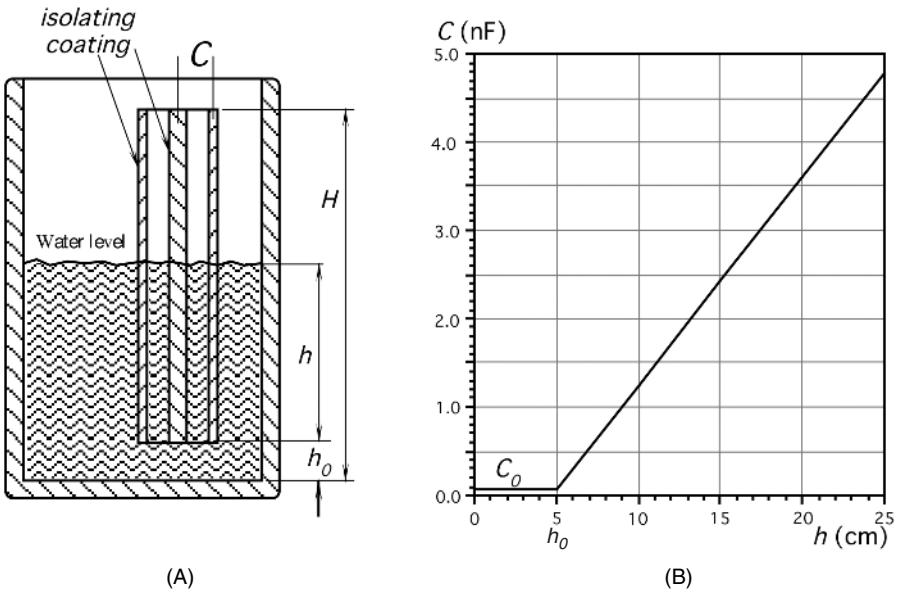


Fig. 3.8. Capacitive water level sensor (A); capacitance as a function of the water level (B).

where h is height of the water-filled portion of the sensor. If the water is at or below the level h_0, the capacitance remains constant

$$C_0 = \frac{2\pi\varepsilon_0}{\ln(b/a)} H. \tag{3.27}$$

Figure 3.8B shows a water level-capacitance dependence.[3] It is a straight line from the level h_0. Because dielectric constant of water is temperature dependent (Fig. 3.7) the capacitive sensor will be combined with a temperature sensor—for instance, a thermistor or resistive temperature detector which monitors water temperature. The appropriate temperature correction may be performed by the electronic signal conditioner.

The slope of the transfer function line depends on the liquid. For instance, if instead of water the sensor measures the level of transformer oil, it is expected to be 22 times less sensitive (see Table A.5).

Another example of a capacitive sensor is a humidity sensor. In such a sensor, a dielectric between the capacitor plates is fabricated of a material that is hygroscopic; that is, it can absorb water molecules and change its dielectric constant accordingly. According to Eq. (3.24), this changes the capacitance that can be measured and related to relative humidity. Figure 3.9 illustrates the dependence between capacitance and relative humidity of such a sensor. The dependence is not linear, but this usually can be taken care of during the signal processing.

[3] The sensor's dimensions are as follows: $a = 10$ mm, $b = 12$ mm, $H = 200$ mm, liquid—water.

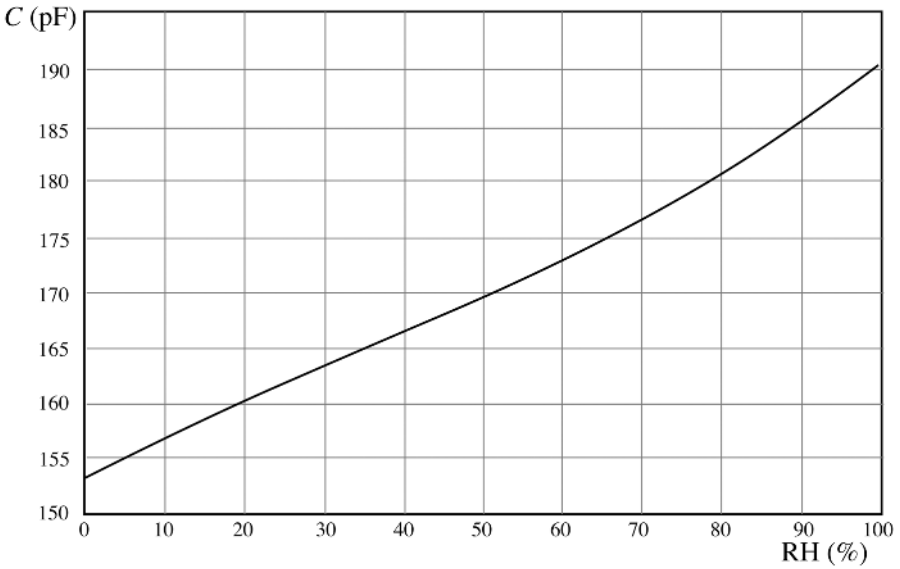


Fig. 3.9. Transfer function of a capacitive relative humidity sensor.

3.3 Magnetism

Magnetic properties were discovered in prehistoric times in certain specimens of an iron ore mineral known as magnetite (Fe_3O_4). It was also discovered that pieces of soft iron that rubbed against a magnetic material acquired the same property of acting as a magnet (i.e., attracting other magnets and pieces of iron). The first comprehensive study of magnetism was made by William Gilbert. His greatest contribution was his conclusion that the Earth acts as a huge magnet. The word "magnetism" comes from the district of Magnesia in Asia Minor, which is one of the places at which the magnetic stones were found.

There is a strong similarity between electricity and magnetism. One manifestation of this is that two electrically charged rods have like and unlike ends, very much in the same way as two magnets have opposite ends. In magnets, these ends are called S (south) and N (north) poles. The like poles repel and the unlike attract. Contrary to electric charges, the magnetic poles always come in pairs. This is proven by breaking magnets into any number of parts. Each part, no matter how small, will have a north pole and a south pole. This suggests that the cause of magnetism is associated with atoms or their arrangements or, more probably, with both.

If we place a magnetic pole in a certain space, that space about the pole appears to have been altered from what it was before. To demonstrate this, bring into that space a piece of iron. Now, it will experience a force that it will not experience if the magnet is removed. This altered space is called a magnetic field. The field is considered to exert a force on any magnetic body brought into the field. If that magnetic body is a small bar magnet or a magnetic needle, the magnetic field will be found to have direction.

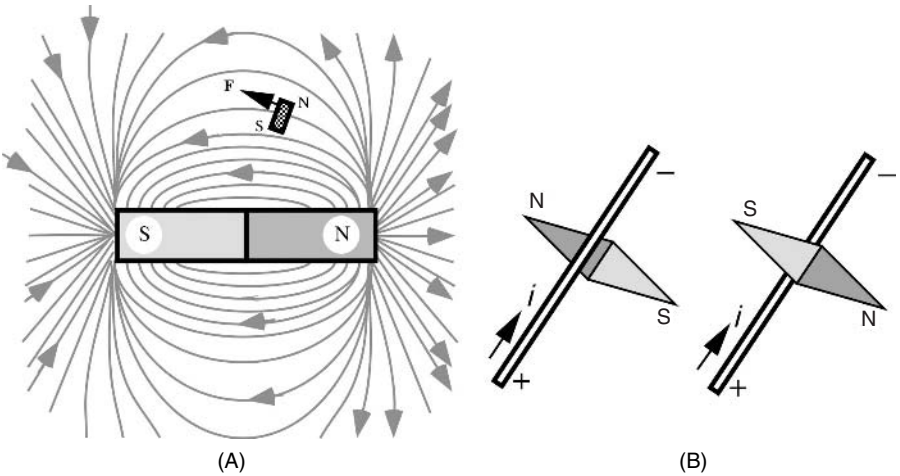


Fig. 3.10. Test magnet in a magnetic field (A); compass needle rotates in accordance with the direction of the electric current (B).

By definition, the direction of this field at any point is given by the direction of the force exerted on a small-unit north pole. Directions of field lines are, by definition, from north to south. Figure 3.10A shows the direction of the field by arrows. A tiny test magnet is attracted in the direction of the force vector **F**. Naturally, approximately the same force but of opposite direction is exerted on the south pole of the test magnet.

The above description of the magnetic field was made for a permanent magnet. However, the magnetic field does not change its nature if it is produced by a different device (e.g., electric current passing through a conductor). It was Hans Christian Oersted, a Danish professor of physics, who in 1820 discovered that a magnetic field could exist where there were no magnets at all. In a series of experiments in which he used an unusually large Voltaic pile (battery) so as to produce a large current, he happened to note that a compass in the near vicinity was behaving oddly. Further investigation showed that the compass needle always oriented itself at right angles to the current-carrying wire and that it reversed its direction if either current was reversed, or the compass was changed from a position below the wire to one above (Fig. 3.10B). Stationary electric charges have no effect on a magnetic compass (in this experiment, a compass needle is used as a tiny test magnet). It was clear that the moving electric charges were the cause of the magnetic field. It can be shown that magnetic field lines around a wire are circular and their direction depends on the direction of electric current (i.e., moving electrons) (Fig. 3.11). Above and below the wire, magnetic field lines are pointed in the opposite direction. That is why the compass needle turns around when it is placed below the wire.

A fundamental property of magnetism is that moving electric charges (electric current) essentially produce a magnetic field. Knowing this, we can explain the nature of a permanent magnet. A simplified model of a magnetic field origination process is shown in Fig. 3.12A. An electron continuously spins in an eddy motion around the

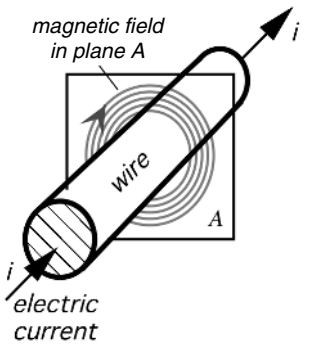


Fig. 3.11. Electric current sets a circular magnetic field around a conductor.

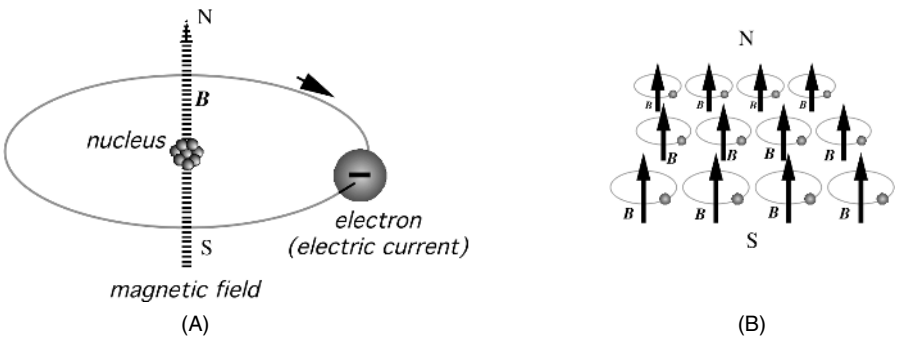


Fig. 3.12. Moving electron sets a magnetic field (A); superposition of field vectors results in a combined magnetic field of a magnet (B).

atom. The electron movement constitutes a circular electric current around the atomic nucleus. That current is a cause for a small magnetic field. In other words, a spinning electron forms a permanent magnet of atomic dimensions. Now, let us imagine that many of such atomic magnets are aligned in an organized fashion (Fig. 3.12B), so that their magnetic fields add up. The process of magnetization then becomes quite obvious: Nothing is added or removed from the material—only the orientation of atoms is made. The atomic magnets may be kept in the aligned position in some materials which have an appropriate chemical composition and a crystalline structure. Such materials are called *ferromagnetics*.

3.3.1 Faraday's Law

Michael Faraday pondered the question, "If an electric current is capable of producing magnetism, is it possible that magnetism can be used to produce electricity?" It took him 9 or 10 years to discover how. If an electric charge is moved across a magnetic field, a deflecting force is acting on that charge. It must be emphasized that it is not important what actually moves—either the charge or the source of the magnetic field. What matters is a relative displacement of those. A discovery that a moving

electric charge can be deflected as a result of its interaction with the magnetic field is a fundamental in electromagnetic theory. Deflected electric charges result in an electric field generation, which, in turn, leads to a voltage difference in a conducting material, thus producing an electric current.

The intensity of a magnetic field at any particular point is defined by a vector **B**, which is tangent to a magnetic field line at that point. For a better visual representation, the number of field lines per unit cross-sectional area (perpendicular to the lines) is proportional to the magnitude of **B**. Where the lines are close together, **B** is large, and where they are far apart, **B** is small.

The flux of magnetic field can be defined as

$$\Phi_B = \oint \mathbf{B}\, ds, \tag{3.28}$$

where the integral is taken over the surface for which $\mathbf{F_B}$ is defined.

To define the magnetic field vector **B**, we use a laboratory procedure where a positive electric charge q_0 is used as a test object. The charge is projected through the magnetic field with velocity **V**. A sideways deflecting force $\mathbf{F_B}$ acts on the charge (Fig. 3.13A). By "sideways," we mean that $\mathbf{F_B}$ is at a right angle to **V**. It is interesting to note that the vector **V** changes its direction while moving through the magnetic field. This results in a spiral rather than parabolic motion of the charge (Fig. 3.13B). The spiral movement is a cause for a magnetoresistive effect which forms a foundation for the magnetoresistive sensors. The deflecting force $\mathbf{F_B}$ is proportional to the charge, velocity, and magnetic field:

$$\mathbf{F_B} = q_0 \mathbf{V} \mathbf{B}. \tag{3.29}$$

The vector $\mathbf{F_B}$ is always at right angles to the plane formed by **V** and **B** and, thus, is always at right angles to **v** and to **B**, that is why it is called a sideways force. The magnitude of magnetic deflecting force according to the rules for vector products is

$$F_B = q_0 \mathrm{v} B \sin \phi, \tag{3.30}$$

where ϕ is the angle between vectors **V** and **B**. The magnetic force vanishes if **V** is parallel to **B**. Equation (3.30) is used for the definition of the magnetic

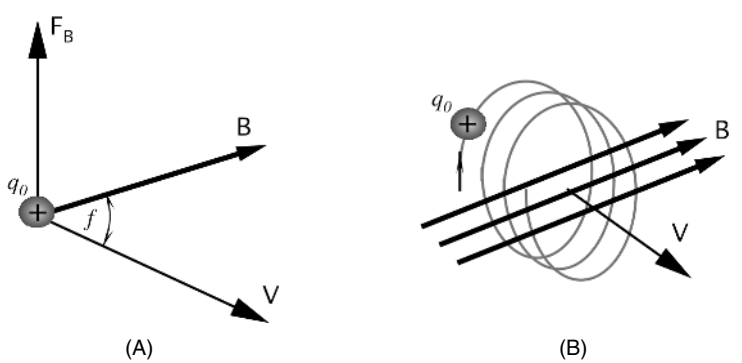


Fig. 3.13. Positive charge projected through a magnetic field is subjected to a sideways force (A); spiral movement of an electric charge in a magnetic field (B).

field in terms of deflected charge, its velocity, and deflecting force. Therefore, the units of B is (Newton/coulomb)/(meter/second)$^{-}$1. In the SI system, it is given the name *tesla* (abbreviated T). Because coulomb/second is an ampere, we have 1 T=1 Newton/(ampere meter). An older unit for B is still in use. It is the gauss: 1 tesla $= 10^4$ gauss.

3.3.2 Solenoid

A practical device for producing a magnetic field is called a *solenoid*. It is a long wire wound in a close-packed helix and carrying a current i. In the following discussion, we assume that the helix is very long compared to its diameter. The solenoid magnetic field is the *vector sum* of the fields set up by all the turns that make up the solenoid.

If a coil (solenoid) has widely spaced turns, the fields tend to cancel between the wires. At points inside the solenoid and reasonably far from the wires, **B** is parallel to the solenoid axis. In the limiting case of adjacent very tightly packed wires (Fig. 3.14A), the solenoid becomes essentially a cylindrical current sheet. If we apply Ampere's law to that current sheet, the magnitude of magnetic field inside the solenoid becomes

$$B = \mu_0 i_0 n, \tag{3.31}$$

where n is the number of turns per unit length and i_0 is the current through the solenoid wire. Although, this formula was derived for an infinitely long solenoid, it holds quite well for actual solenoids for internal points near the center of the solenoid. It should be noted that B does not depend on the diameter or the length of the solenoid and that B is constant over the solenoid cross section. Because the solenoid's diameter is not a part of the equation, multiple layers of winding can be used to produce a magnetic field of higher strength. It should be noted that the magnetic field outside of a solenoid is weaker than that of the inside.

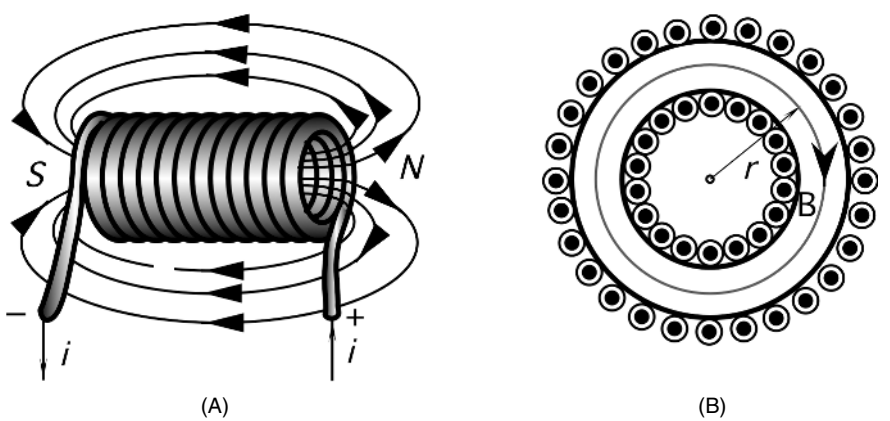


Fig. 3.14. Solenoid (A) and toroid (B).

3.3.3 Toroid

Another useful device that can produce a magnetic field is a toroid (Fig. 3.14B), which we can describe as a solenoid bent into the shape of a doughnut. A calculation of the magnetic field inside the toroid gives the following relationship:

$$B = \frac{\mu_0}{2\pi} \frac{i_0 N}{r}, \tag{3.32}$$

where N is the total number of turns and r is the radius of the inner circular line where the magnetic field is calculated. In contrast to a solenoid, B is not constant over the cross section of a toroid. Also, for an ideal case, the magnetic field is equal to zero outside a toroid.

The density of a magnetic field, or the number of magnetic lines passing through a given surface, is defined as the magnetic flux $\mathbf{\Phi_B}$ for that surface:

$$\mathbf{\Phi_B} = \int \mathbf{B}\, dS. \tag{3.33}$$

The integral is taken over the surface, and if the magnetic field is constant and is everywhere at a right angle to the surface, the solution of the integral is very simple: $\mathbf{\Phi_B} = \mathbf{B}A$, where A is the surface area. Flux, or flow of the magnetic field, is analogous to the flux of electric field. The SI unit for magnetic flux, as follows from the above, is tesla meter^2, to which is named *weber*. It is abbreviated as Wb:

$$1\text{Wb} = 1\text{T m}^2. \tag{3.34}$$

3.3.4 Permanent Magnets

Permanent magnets are useful components for fabricating magnetic sensors for the detection of motion, displacement, position, and so forth. To select the magnet for any particular application, the following characteristics should be considered:

- Residual inductance (B) in gauss—how strong the magnet is?
- Coercive force (H) in oersteds—how well will the magnet resist external demagnetization forces?
- Maximum energy product, MEP, (BH) is gauss oersteds times 10^6. A strong magnet that is also very resistant to demagnetization forces has a high MEP. Magnets with a higher MEP are better, stronger, and more expensive.
- The temperature coefficient in %/°C shows how much B changes with temperature.

Magnets are produced from special alloys (see Table A.6). Examples are *rare earth* (e.g., samarium)–*cobalt* alloys. These are the best magnets; however, they are too hard for machining and must be ground if shaping is required. Their maximum MEP is about 16×10^6. Another popular alloy is *Alnico*, which contains aluminum, nickel, cobalt, iron, and some additives. These magnets can be cast or sintered by

pressing metal powders in a die and heating them. Sintered Alnico is well suited to mass production. *Ceramic magnets* contain barium or strontium ferrite (or another element from that group) in a matrix of a ceramic material that is compacted and sintered. They are poor conductors of heat and electricity, are chemically inert, and have a high value of H. Another alloy for the magnet fabrication is *Cunife*, which contains copper, nickel, and iron. It can be stamped, swaged, drawn, or rolled into final shape. Its MEP is about 1.4×10^6. *Iron–chromium magnets* are soft enough to undergo machining before the final aging treatment hardens them. Their maximum MEP is 5.25×10^6. *Plastic and rubber* magnets consist of barium or strontium ferrite in a plastic matrix material. They are very inexpensive and can be fabricated in many shapes. Their maximum MEP is about 1.2×10^6.

3.4 Induction

In 1831, Michael Faraday in England and Joseph Henry in the United States discovered one of the most fundamental effects of electromagnetism: an ability of a varying magnetic field to induce electric current in a wire. It is not important how the field is produced—either by a permanent magnet or by a solenoid—the effect is the same. Electric current is generated as long as the magnetic field *changes*. A stationary field produces no current. Faraday's law of induction says that the induced voltage, or electromotive force (*e.m.f.*), is equal to the rate at which the magnetic flux through the circuit changes. If the rate of change is in webers per second, the e.m.f. (e) will be in volts:

$$e = -\frac{d\Phi_B}{dt}. \tag{3.35}$$

The minus sign is an indication of the direction of the induced e.m.f. If varying magnetic flux is applied to a solenoid, the e.m.f. appears in every turn and all of these e.m.f.'s must be added. If a solenoid, or other coil, is wound in such a manner so that each turn has the same cross-sectional area, the flux through each turn will be the same, and then the induced voltage is

$$V = -N\frac{d\Phi_B}{dt}, \tag{3.36}$$

where N is the number of turns. This equation may be rewritten in a form which is of interest to a sensor designer or an application engineer:

$$V = -N\frac{d(BA)}{dt}. \tag{3.37}$$

The equation means that the voltage in a pickup circuit can be produced by either changing the amplitude of the magnetic field (B) or area of the circuit (A). Thus, induced voltage depends on the following:

- Moving the source of the magnetic field (magnet, coil, wire, etc.)
- Varying the current in the coil or wire which produces the magnetic field

- Changing the orientation of the magnetic source with respect to the pickup circuit
- Changing the geometry of a pickup circuit, (e.g., by stretching it or squeezing, or changing the number of turns in a coil)

If an electric current passes through a coil which is situated in close proximity with another coil, according to Faraday's law, the e.m.f. in a second coil will appear. However, the magnetic field penetrates not only the second coil, but the first coil as well. Thus, the magnetic field sets the e.m.f. in the same coil where it is originated. This is called *self-induction* and the resulting voltage is called a *self-induced e.m.f.* Faraday's law for a central portion of a solenoid is

$$v = -\frac{d(n\Phi_B)}{dt}. \tag{3.38}$$

The number in parentheses is called the flux linkage and is an important characteristic of the device. For a simple coil with no magnetic material in the vicinity, this value is proportional to the current through the coil:

$$n\Phi_B = Li, \tag{3.39}$$

where L is a proportionality constant, which is called the *inductance* of the coil. Then, Eq. (3.38) can be rewritten as

$$v = -\frac{d(n\Phi_B)}{dt} = -L\frac{di}{dt}. \tag{3.40}$$

From this equation, we can define inductance as

$$L = -\frac{v}{di/dt} \tag{3.41}$$

If no magnetic material is introduced in the vicinity of an *inductor* (a device possessing inductance), the value defined by Eq. (3.41) depends only on the geometry of the device. The SI unit for inductance is the volt second/ampere, which was named after American physicist Joseph Henry (1797–1878): 1 henry = 1 volt second/ampere. The abbreviation for henry is H.

Several conclusions can be drawn from Eq. (3.41):

- Induced voltage is proportional to the rate of change in current through the inductor,
- Voltage is essentially zero for direct current (dc).
- Voltage increases linearly with the current rate of change.
- Voltage polarity is different for increased and decreased currents flowing in the same direction.
- Induced voltage is always in the direction which opposes the change in current.

Like capacitance, inductance can be calculated from geometrical factors. For a closely packed coil, it is

$$L = \frac{n\Phi_B}{i}. \tag{3.42}$$

If n is the number of turns per unit length, the number of flux linkages in the length, l, is

$$N\Phi_B = (nl)(BA), \tag{3.43}$$

where A is the cross-sectional area of the coil. For the solenoid, $B = \mu_0 ni$, and the inductance is

$$L = \frac{N\Phi_B}{i} = \mu_0 n^2 lA. \tag{3.44}$$

It should be noted that lA is the volume of a solenoid, Thus, having the same number of turns and changing the coil geometry, its inductance may be modulated (altered).

When connected into an electronic circuit, inductance may be represented as a "complex resistance":

$$\frac{V}{i} = j\omega L, \tag{3.45}$$

where $j = \sqrt{-1}$ and i is a sinusoidal current having a frequency of $\omega = 2\pi f$, meaning that the complex resistance of an inductor increases at higher frequencies. This is called Ohm's law for an inductor. Complex notation indicates that current lags behind voltage by 90°.

If two coils are brought in the vicinity of one another, one coil induces e.m.f., v_2 in the second coil:

$$v_2 = -M_{21}\frac{di_1}{dt}, \tag{3.46}$$

where M_{21} is the coefficient of mutual inductance between two coils. The calculation of mutual inductance is not a simple exercise, and in many practical cases, it can be easier performed experimentally. Nevertheless, for some relatively simple combinations, mutual inductances have been calculated. For a coil (having N turns) which is placed around a long solenoid (Fig. 3.15A), having n turns per unit length, mutual inductance is

$$M = \mu_0 \pi R^2 nN. \tag{3.47}$$

For a coil placed around a toroid (Fig. 3.15B), mutual inductance is defined by the numbers of turns, N_1 and N_2:

$$M = \frac{\mu_0 N_1 N_2 h}{2\pi} \ln\left(\frac{b}{a}\right) \tag{3.48}$$

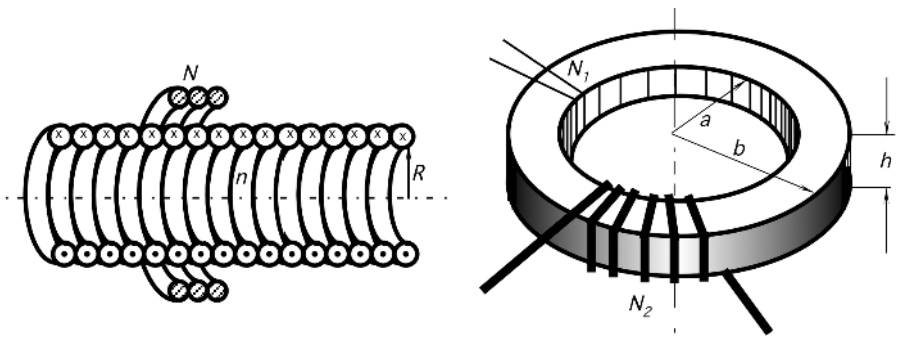


Fig. 3.15. Mutual inductances in solenoids (A) and in a toroid (B).

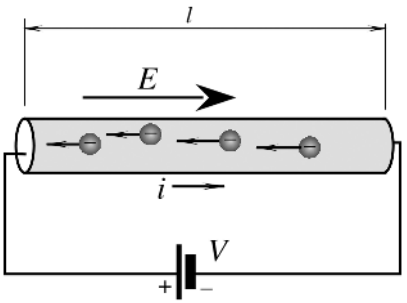


Fig. 3.16. Voltage across a material sets the electric current.

3.5 Resistance

In any material, electrons move randomly like gas in a closed container. There is no preferred direction and an average concentration of electrons in any part of the material is uniform (assuming that the material is homogeneous). Let us take a bar of an arbitrary material. The length of the bar is l. When the ends of the bar are connected to the battery having voltage V (Fig. 3.16), an electric field **E** will be setup within the material. It is easy to determine the strength of the electric field:

$$E = \frac{V}{l} \tag{3.49}$$

For instance, if the bar has a length of 1 m and the battery delivers 1.5 V, the electric field has a strength of 1.5 V/m. The field acts on free electrons and sets them in motion against the direction of the field. Thus, the electric current starts flowing through the material. We can imagine a cross section of the material through which passes electric charge q. The rate of the electric charge flowing (unit of charge per unit of time) is called the electric current:

$$i = \frac{dq}{dt} \tag{3.50}$$

The SI unit of current is ampere (A): 1 A = 1 coulomb/sec. In SI units, ampere is defined as the electric current maintained in two infinitely long parallel wires separated by 1 m in free space, which produce a force between the two wires (due to their magnetic field) of 2×10^{-7} N for each meter of length. An ampere is quite strong electric current. In sensor technologies, generally much smaller currents are used; therefore, submultiples of A are often employed:

1 milliampere (mA): 10^{-3} A
1 microampere (μA): 10^{-6} A
1 nanoampere (nA): 10^{-9} A
1 picoampere (pA): 10^{-12} A
1 femtoampere (fA): 10^{-15} A

Regardless of the cross section of the material, whether it is homogeneous or not, the electric current through any cross section is always the same for a given electric field. It is similar to water flow through a combination of serially connected pipes of

different diameters; the rate of flow is the same throughout of the pipe combination. The water flows faster in the narrow sections and slower in the wide section, but the amount of water passing through any cross section per unit of time is constant. The reason for that is very simple: water in the pipes is neither drained out nor created. The same reason applies to electric current. One of the fundamental laws of physics is the law of conservation of charge. Under steady-state conditions, charge in a material is neither created nor destroyed. *Whatever comes in must go out*. In this section, we do not consider any charge storages (capacitors), and all materials we discuss are said have pure *resistive* properties.

The mechanism of electrical conduction in a simplified form may be described as follows. A conducting material, say copper wire, can be modeled as a semi-rigid springlike periodic lattice of positive copper ions. They are coupled together by strong electromagnetic forces. Each copper atom has one conduction electron which is free to move about the lattice. When electric field $\mathbf{E}$ is established within the conductor, the force $-e\mathbf{E}$ acts on each electron (e is the electron charge). The electron accelerates under the force and moves. However, the movement is very short, as the electron collides with the neighboring copper atoms, which constantly vibrate with an intensity determined by the material temperature. The electron transfers kinetic energy to the lattice and is often captured by the positive ion. It frees another electron, which keeps moving in the electric field until, in turn, it collides with the next portion of the lattice. The average time between collisions is designated as τ. It depends on the material type, structure, and impurities. For instance, at room temperature, a conduction electron in pure copper moves between collisions for an average distance of 0.04 μm, with $\tau = 2.5 \times 10^{-14}$s. In effect, electrons which flow into the material near the negative side of the battery are not the same which outflow to the positive terminal. However, the constant drift or flow of electrons is maintained throughout the material. Collisions of electrons with the material atoms further add to the atomic agitation and, subsequently, raise the material temperature. This is why passing of electric current through a resistive material results in the so-called Joule heat liberation.

It was arbitrarily decided to define the direction of current flow along with the direction of the electric field (i.e., in the *opposite direction* of the electronic flow). Hence, the electric current flows from the positive to negative terminal of the battery while electrons actually move in the opposite direction.

3.5.1 Specific Resistivity

If we fabricate two geometrically identical rods from different materials, say from copper and glass, and apply to them the same voltage, the resulting currents will be quite different. A material may be characterized by its ability to pass electric current. It is called *resistivity* and material is said to have electrical *resistance* which is defined by Ohm's law:

$$R = \frac{V}{i}. \tag{3.51}$$

For pure resistance (no inductance or capacitance), voltage and current are in-phase with each other, meaning that they are changing simultaneously.

Any material has electric resistivity[4] and therefore is called a *resistor*. The SI unit of resistance is 1 ohm (Ω) = 1 volt/1 ampere. Other multiples and submultiples of Ω are as follows:

1 milliohm (mΩ): 10^{-3} Ω
1 kilohm (kΩ): 10^{3} Ω
1 megohm (MΩ): 10^{6} Ω
1 gigohm (GΩ): 10^{9} Ω
1 terohm (TΩ): 10^{12} Ω

If we compare electric current with water flow, pressure across the pipe line (Pascal) is analogous to voltage (V) across the resistor, electric current (C/s) is analogous to water flow (L/s), and electric resistance (Ω) corresponds to water flow resistance in the pipe. It is clear that resistance to water flow is lower when the pipe is short, wide, and empty. When the pipe has, for instance, a filter installed in it, resistance to water flow will be higher. Similarly, coronary blood flow may be restricted by cholesterol deposits on the inner lining of blood vessels. Flow resistance is increased and arterial blood pressure is not sufficient to provide the necessary blood supply rate for normal functioning of the heart. This may result in a heart attack. The basic laws that govern the electric circuit designs are called Kirchhoff's laws, after the German physicist Gustav Robert Kirchhoff (1824–1887). These laws were originally devised for the plumbing networks, which, as we have seen, are similar to electric networks.

Resistance is a characteristic of a device. It depends on both the material and the geometry of the resistor. Material itself can be characterized by a *specific resistivity*, ρ, which is defined as

$$\rho = \frac{E}{j}, \tag{3.52}$$

where j is current density: $j = i/a$ (a is the area of the material cross section). The SI unit of resistivity is Ω m. Resistivities of some materials are given in the Appendix (Table A.7). Quite often, a reciprocal quantity is used which is called *conductivity*: $\sigma = 1/\rho$.

The resistivity of a material can be expressed through mean time between collisions, τ, the electronic charge, e, the mass of electron, m, and a number of conduction electrons per unit volume, n:

$$\rho = \frac{m}{ne^2\tau}. \tag{3.53}$$

To find the resistance of a conductor, the following formula may be used:

$$R = \rho\frac{l}{a}, \tag{3.54}$$

where a is the cross sectional area and l is the length of the conductor.

[4] Excluding superconductors, which are beyond the scope of this book.

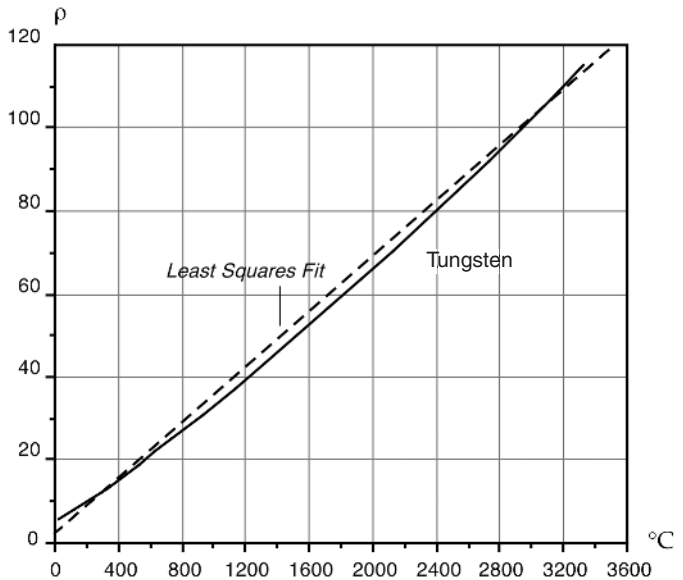


Fig. 3.17. Specific resistivity of tungsten as a function of temperature.

3.5.2 Temperature Sensitivity

The conductivity of a material changes with temperature, t, and in a relatively narrow range, it may be expressed by α, which is temperature coefficient of resistance (TCR):

$$\rho = \rho_0[1 + \alpha(t - t_0)] \tag{3.55}$$

where ρ_0 is resistivity at the reference temperature t_0 (commonly either 0°C or 25°C). In a broader range, resistivity is a nonlinear function of temperature.

For nonprecision applications over a broad temperature range, the resistivity of tungsten, as shown in Fig. 3.17 may be modeled by a best-fit straight line with $\alpha = 0.0058\ \mathrm{C}^{-1}$. However, this number will not be accurate at lower temperatures. For instance, near 25°C the slope of ρ is about 20% smaller: $\alpha = 0.0045\ \mathrm{C}^{-1}$. When better accuracy is required, formula (3.55) should not be employed. Instead, higher-order polynomials may be useful for modeling the resistivity. For instance, over a broader temperature range, tungsten resistivity may be found from the second-order equation

$$\rho = 4.45 + 0.0269t + 1.914 \times 10^{-6} t^2, \tag{3.56}$$

where t is the temperature (in °C) and ρ is in Ω m.

Metals have positive temperature coefficients (PTCs) α, whereas many semiconductors and oxides have negative temperature coefficients of resistance (NTCs). It is usually desirable to have very low TCRs in resistors used in electronic circuits. On the other hand, a strong temperature coefficient of resistivity allows us to fabricate a temperature sensor, known as a *thermistor* (a contraction of the words *thermal* and

resistor) and the so-called resistive temperature detector (RTD).[5] The most popular RTD is a platinum (Pt) sensor which operates over a broad temperature range from about −200°C to over 600°C. The resistance of a Pt RTD is shown in Fig. 3.18. For a calibrating resistance R_0 at 0°C, the best-fit straight line is given by

$$R = R_0(1.00 + 36.79 \times 10^{-4}t), \tag{3.57}$$

where t is the temperature in °C and R is in Ω. The multiple at temperature (T) is the sensor's sensitivity (a slope), which may be expressed as +0.3679%/°C.

There is a slight nonlinearity of the resistance curve which, if not corrected, may lead to an appreciable error. A better approximation of Pt resistance is a second-order polynomial which gives an accuracy better than 0.01°C:

$$R = R_0(1 + 39.08 \times 10^{-4}t - 5.8 \times 10^{-7}t^2)\Omega. \tag{3.58}$$

It should be noted, however, that the coefficients in Eqs. (3.57) and (3.58) somewhat depend on the material purity and manufacturing technologies. To compare accuracies of the linear and the second-order models of the platinum thermometer, consider the following example. If a Pt RTD sensor at 0°C has resistivity $R_0 = 100\Omega$, at +150°C, the linear approximation gives

$$R = 100[1.0036 + 36.79 \times 10^{-4}(150)] = 155.55\Omega,$$

whereas for the second-order approximation (Eq. 3.58)

$$R = 100[1 + 39.08 \times 10^{-4}150 - 5.8 \times 10^{-7}(150)^2] = 157.32\Omega.$$

The difference between the two is 1.76 Ω. This is equivalent to an error of −4.8°C at +150°C.

Thermistors are resistors with large either negative (NTC) or positive (PTC) temperature coefficients. The thermistors are ceramic semiconductors commonly made of oxides of one or more of the following metals: nickel, manganese, cobalt, titanium, iron. Oxides of other metals are occasionally used. Resistances vary from a fraction of an ohm to many megohms. Thermistors can be produced in the form of disks, droplets, tubes, flakes, or thin films deposited on ceramic substrates. Recent progress in thick-film technology allows us to print the thermistor on ceramic substrates.

The NTC thermistors often are fabricated in the form of beads. Usually, bead thermistors have platinum alloy lead wires which are sintered into the ceramic body. Platinum is a convenient choice for the thermistor wires because it combines a relatively low electrical resistance with a relatively high thermal resistance. During the fabrication process, a small portion of mixed metal oxides is placed with a suitable binder onto a pair of platinum alloy wires, which are under slight tension. After the mixture has been allowed to set, the beads are sintered in a tubular furnace. The metal oxides shrink around the platinum lead wires and form intimate electrical bonds. The beads may be left bare or they may be given organic or glass coatings.

[5] See Section 16.1 of Chapter 16

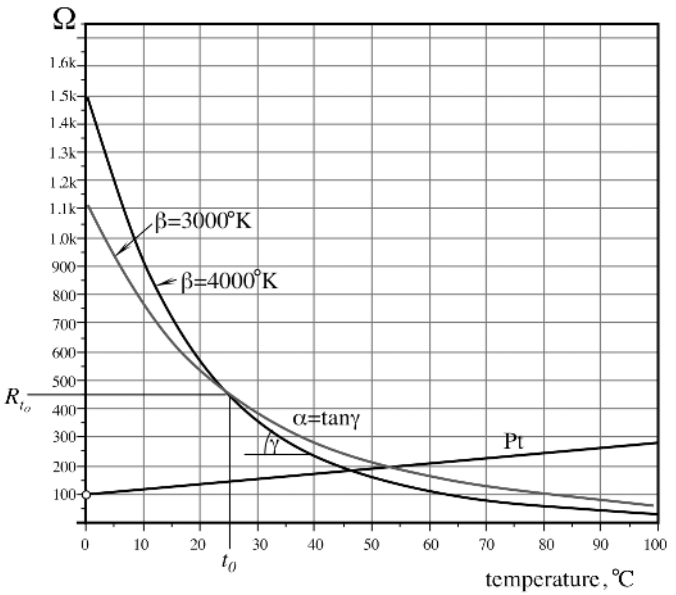


Fig. 3.18. Resistance–temperature characteristics for two thermistors and Pt RTD ($R_0 = 1k$); thermistors are calibrated at $t_0 = 25°\text{C}$ and RTD at 0°C.

Thermistors possess nonlinear temperature–resistance characteristics (Fig. 3.18), which are generally approximated by one of several different equations. The most popular of them is the exponential form

$$R_t = R_{t_0} \mathrm{e}^{\beta(1/T - 1/To)}, \tag{3.59}$$

where T_0 is the calibrating temperature in kelvin, R_{t_0} is the resistance at the calibrating temperature, and β is a material's characteristic temperature. All temperatures and β are in kelvin. Commonly, β ranges between 3000 and 5000 K, and for a relatively narrow temperature range, it can be considered temperature independent, which makes Eq. (3.59) a reasonably good approximation. When a higher accuracy is required, a polynomial approximation is generally employed. Figure 3.18 shows the resistance–temperature dependence of thermistors having $\beta = 3000$ and 4000K and that for the platinum RTD. The temperature characteristic of platinum is substantially less sensitive and more linear with a positive slope, whereas thermistors are nonlinear with a high sensitivity and a negative slope.

Traditionally, thermistors are specified at temperature of $t_0 = 25°\text{C}$ ($T_0 = 298.15°\text{K}$), whereas RTDs are specified at $t_0 = 0°\text{C}$ ($T_0 = 273.15°\text{K}$).

3.5.3 Strain Sensitivity

Usually, electrical resistance changes when the material is mechanically deformed. This is called the *piezoresistive effect*. In some cases, the effect is a source of error. On

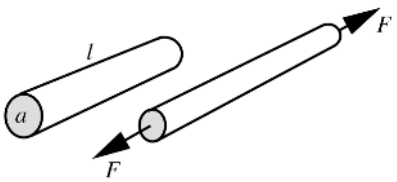


Fig. 3.19. Strain changes the geometry of a conductor and its resistance.

the other hand, it is successfully employed in sensors which are responsive to stress, σ:

$$\sigma = \frac{F}{a} = E\frac{dl}{l}, \tag{3.60}$$

where E is Young's modulus of the material and F is the applied force. In this equation, the ratio $dl/l = e$ is called *strain*, which is a normalized deformation of the material.

Figure 3.19 shows a cylindrical conductor (wire) stretched by applied force F. The volume v of the material stays constant while the length increases and the cross sectional area becomes smaller. As a result, Eq. (3.54) can be rewritten as

$$R = \frac{\rho}{v} l_2. \tag{3.61}$$

After differentiating, we can define sensitivity of resistance with respect to wire elongation:

$$\frac{dR}{dl} = 2\frac{\rho}{v} l. \tag{3.62}$$

It follows from this equation that the sensitivity becomes higher for the longer and thinner wires with a high specific resistance. Normalized incremental resistance of the strained wire is a linear function of strain, e, and it can be expressed as

$$\frac{dR}{R} = S_e e, \tag{3.63}$$

where S_e is known as the *gauge factor* or *sensitivity* of the strain gauge element. For metallic wires, it ranges from 2 to 6. It is much higher for semiconductor gauges; it is between 40 and 200.

Early strain gauges were metal filaments. The gauge elements were formed on a backing film of electrically isolating material. Today, they are manufactured from constantan (a copper/nickel alloy) foil or single-crystal semiconductor materials (silicon with boron impurities). The gauge pattern is formed either by mechanical cutting or photochemical etching. When a semiconductor material is stressed, its resistivity changes depending on the type of the material and the doping dose (see Section 9.1 of Chapter 9). However, the strain sensitivity in semiconductors is temperature dependent which requires a proper compensation when used over a broad temperature range.

3.5.4 Moisture Sensitivity

By selecting material for a resistor, one can control its specific resistivity and susceptibility of such to the environmental factors. One of the factors is the amount of

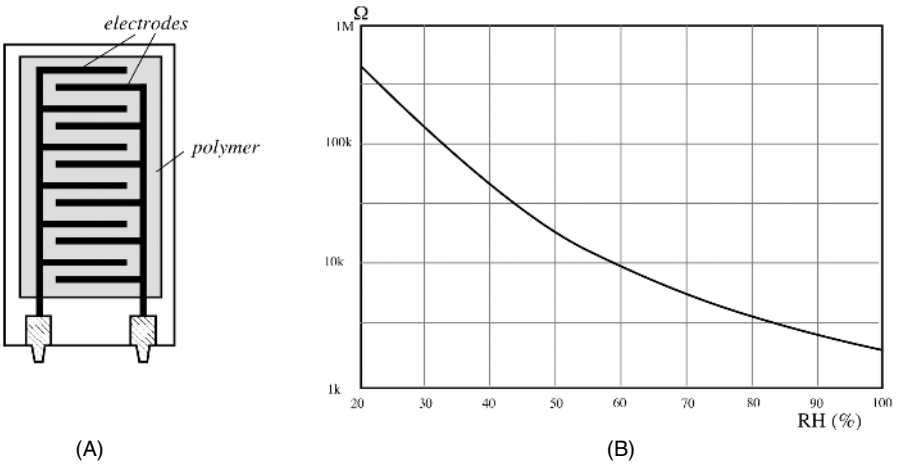


Fig. 3.20. Hygristor design (A) and its transfer function (B).

moisture that can be absorbed by the resistor. A moisture-dependent resistor can be fabricated of hygroscopic material whose specific resistivity is strongly influenced by the concentration of the absorbed water molecules. This is the basis for the resistive humidity sensors, which are called *hygristors*. A typical design of a hygristor is comprised of a substrate that has two silkscreen-printed conductive interdigitized electrodes (Fig. 3.20A) which are covered and thus electrically interconnected by hygroscopic semiconductive gel which forms a matrix to hold the conductive particles. The gel [2] is typically fabricated of hydroxyethylcellulose, nonylphenylpolyethylene glycol ether, and other (exotically sounding names for an electrical engineer!) organic materials with the addition of carbon powder. The gel is thoroughly milled to produce a smooth mixture. Another type of hygristor is fabricated of lithium chloride (LiCl) film and a binder. The sensor substrates are dipped into the milled gel at controlled rates to coat the gel on the space between the electrodes. The coated substrates are cured under controlled temperature and humidity. Resistance changes with humidity in a nonlinear way (Fig. 3.20B), which can be taken into account during the calibration and data processing.

3.6 Piezoelectric Effect

The piezoelectric effect is the generation of electric charge by a crystalline material upon subjecting it to stress. The effect exists in natural crystals, such as quartz (chemical formula SiO_2), and poled (artificially polarized) man-made ceramics and some polymers, such as polyvinylidene flouride. It is said that piezoelectric material possess ferroelectric properties. The name was given by an analogy with ferromagnetic properties. The word *piezo* comes from the Greek *piezen*, meaning to press. The Curie brothers discovered the piezoelectric effect in quartz in 1880, but very little

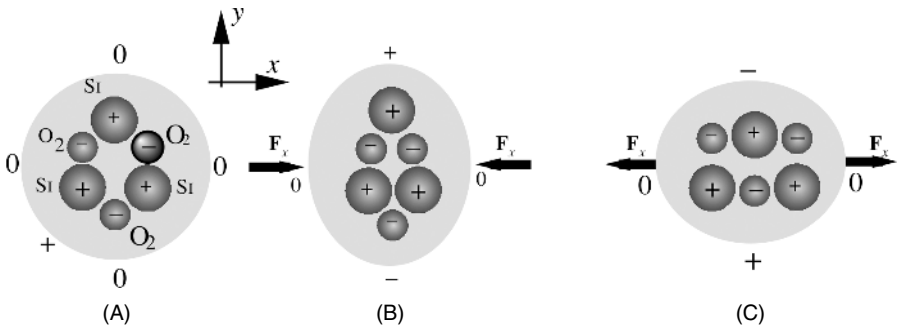


Fig. 3.21. Piezoelectric effect in a quartz crystal.

practical use was made until 1917, when another Frenchman, Professor P. Langevin used x-cut plates of quartz to generate and detect sound waves in water. His work led to the development of sonar.

A simplified, yet quite explanatory model of the piezoelectric effect was proposed in 1927 by A. Meissner [3]. A quartz crystal is modeled as a helix (Fig. 3.21A) with one silicon, Si, and two oxygen, O_2, atoms alternating around the helix. A quartz crystal is cut along its axes x, y, and z; thus, Fig. 3.21A is a view along the z-axis. In a single-crystal cell, there are three silicon atoms and six oxygen atoms. Oxygen is being lumped in pairs. Each silicon atom carries four positive charges and a pair of oxygen atoms carries four negative charges (two per atom). Therefore, a quartz cell is electrically neutral under the no-stress conditions. When an external force, F_x, is applied along the x-axis, the hexagonal lattice becomes deformed. Figure 3.21B shows a compressing force which shifts atoms in a crystal in such a manner that a positive charge is built up at the silicon atom side and a negative charge at the oxygen pair side. Thus, the crystal develops an electric charge along the y-axis. If the crystal is stretched along the x-axis (Fig. 3.21C), a charge of opposite polarity is built along the y-axis, which is a result of a different deformation. This simple model illustrates that crystalline material can develop electric charge on its surface in response to a mechanical deformation. A similar explanation may be applied to the pyroelectric effect, which is covered in Section 3.6.

To pick up an electric charge, conductive electrodes must be applied to the crystal at the opposite sides of the cut (Fig. 3.22). As a result, a piezoelectric sensor becomes a capacitor with a dielectric material which is a piezoelectric crystal. The dielectric acts as a generator of electric charge, resulting in voltage V across the capacitor. Although charge in a crystalline dielectric is formed at the location of an acting force, metal electrodes equalize charges along the surface, making the capacitor not selectively sensitive. However, if electrodes are formed with a complex pattern, it is possible to determine the exact location of the applied force by measuring the response from a selected electrode.

The piezoelectric effect is a reversible physical phenomenon. That means that applying voltage across the crystal produces mechanical strain. By placing several electrodes on the crystal, it is possible to use one pair of electrodes to deliver voltage to

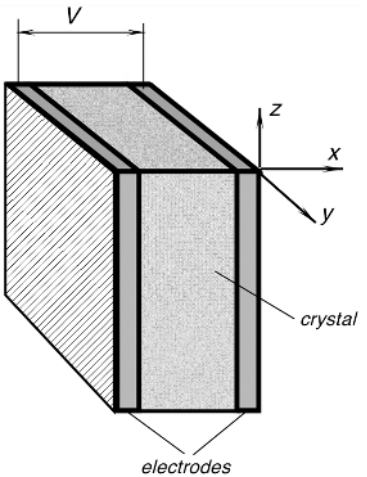


Fig. 3.22. Piezoelectric sensor is formed by applying electrodes to a poled crystalline material.

the crystal and the other pair of electrodes to pick up charge resulting from developed strain. This method is used quite extensively in various piezoelectric transducers.

The magnitude of the piezoelectric effect in a simplified form can be represented by the vector of polarization [4]:

$$\mathbf{P} = \mathbf{P}_{xx} + \mathbf{P}_{yy} + \mathbf{P}_{zz}, \tag{3.64}$$

where x, y, and z refer to a conventional ortogonal system related to the crystal axes. In terms of axial stress, σ, we can write[6]

$$\begin{aligned} \mathbf{P}_{xx} &= d_{11}\boldsymbol{\sigma}_{xx} + d_{12}\boldsymbol{\sigma}_{yy} + d_{13}\boldsymbol{\sigma}_{zz}, \\ \mathbf{P}_{yy} &= d_{21}\boldsymbol{\sigma}_{xx} + d_{22}\boldsymbol{\sigma}_{yy} + d_{23}\boldsymbol{\sigma}_{zz}, \\ \mathbf{P}_{zz} &= d_{31}\boldsymbol{\sigma}_{xx} + d_{32}\boldsymbol{\sigma}_{yy} + d_{33}\boldsymbol{\sigma}_{zz}, \end{aligned} \tag{3.65}$$

where constants d_{mn} are the piezoelectric coefficients along the orthogonal axes of the crystal cut. Dimensions of these coefficients are C/N (coulomb/newton) (i.e., charge unit per unit force).

For the convenience of computation, two additional units have been introduced. The first is a g coefficient which is defined by a division of corresponding d_{mn} coefficients by the absolute dielectric constant

$$g_{mn} = \frac{d_{mn}}{\varepsilon_0 \varepsilon_{mn}}. \tag{3.66}$$

This coefficient represents a voltage gradient (electric field) generated by the crystal per unit applied pressure; its dimension is

$$\frac{\mathrm{V}}{\mathrm{m}} \Big/ \frac{\mathrm{N}}{\mathrm{m}^2}.$$

[6] The complete set of coefficients also includes shear stress and the corresponding d-coefficients.

Another coefficient, designated h, is obtained by multiplying the g coefficients by the corresponding Young's moduli for the corresponding crystal axes. Dimension of the h coefficient is

$$\frac{\mathrm{V}}{\mathrm{m}} \Big/ \frac{\mathrm{m}}{\mathrm{m}}.$$

Piezoelectric crystals are direct converters of mechanical energy into electrical. The efficiency of the conversion can be determined from the so-called *coupling coefficients* k_{mn}:

$$k_{mn} = \sqrt{d_{mn} h_{mn}}. \tag{3.67}$$

The k coefficient is an important characteristic for applications where energy efficiency is of a prime importance, like in acoustics and ultrasonics.

The charge generated by the piezoelectric crystal is proportional to applied force, for instance, in the x direction the charge is

$$Q_x = d_{11} F_x. \tag{3.68}$$

Because a crystal with deposited electrodes forms a capacitor having capacitance C, the voltage, V, which develops across between the electrodes is

$$V = \frac{Q_x}{C} = \frac{d_{11}}{C} F_x. \tag{3.69}$$

In turn, the capacitance can be represented [see Eq. 3.23] through the electrode surface area,[7] a, and the crystal thickness, l:

$$C = \kappa \varepsilon_0 \frac{a}{l}. \tag{3.70}$$

where ε_0 is permitivity constant and κ is dielectric constant. Then, the output voltage is

$$V = \frac{d_{11}}{C} F_x = \frac{d_{11}}{\kappa \varepsilon_0 a} F x. \tag{3.71}$$

The manufacturing of ceramic PZT sensors begins with high purity metal oxides (lead oxide, zirconium oxide, titanium oxide, etc.) in the form of fine powders having various colors. The powders are milled to a specific fineness and mixed thoroughly in chemically correct proportions. In a process called "calcining," the mixtures are then exposed to an elevated temperature, allowing the ingredients to react to form a powder, each grain of which has a chemical composition close to the desired final composition. At this stage, however, the grain does not yet have the desired crystalline structure.

The next step is to mix the calcined powder with solid and/or liquid organic binders (intended to burn out during firing) and mechanically form the mixture into a "cake" which closely approximates a shape of the final sensing element. To form the "cakes" of desired shapes, several methods can be used. Among them are pressing (under force of a hydraulic powered piston), casting (pouring viscous liquid into molds and

[7] Not the crystal area. Piezo-induced charge can be collected only over the area covered by the electrode.

allowing to dry), extrusion (pressing the mixture through a die, or a pair of rolls to form thin sheets), and tape casting (pulling viscous liquid onto a smooth moving belt).

After the "cakes" have been formed, they are placed into a kiln and exposed to a very carefully controlled temperature profile. After burning out of organic binders, the material shrinks by about 15%. The "cakes" are heated to a red glow and maintained at that state for some time, which is called the "soak time," during which the final chemical reaction occurs. The crystalline structure is formed when the material is cooled down. Depending on the material, the entire firing may take 24 h. When the material is cold, contact electrodes are applied to its surface. This can be done by several methods. The most common of them are a fired-on silver (a silkscreening of silver-glass mixture and refiring), an electroless plating (a chemical deposition in a special bath), and a sputtering (an exposure to metal vapor in a partial vacuum).

Crystallities (crystal cells) in the material can be considered electric dipoles. In some materials, like quartz, these cells are naturally oriented along the crystal axes, thus giving the material sensitivity to stress. In other materials, the dipoles are randomly oriented and the materials need to be "poled" to possess piezoelectric properties. To give a crystalline material piezoelectric properties, several poling techniques can be used. The most popular poling process is a thermal poling, which includes the following steps:

1. A crystalline material (ceramic or polymer film) which has randomly oriented dipoles (Fig. 3.23A) is warmed up slightly below its Curie temperature. In some cases (for a PVDF film), the material is stressed. A high temperature results in stronger agitation of dipoles and permits us to more easily orient them in a desirable direction.
2. Material is placed in strong electric field, **E** (Fig. 3.23B) where dipoles align along the field lines. The alignment is not total. Many dipoles deviate from the filed direction quite strongly; however, statistically predominant orientation of the dipoles is maintained.
3. The material is cooled down while the electric field across its thickness is maintained.
4. The electric field is removed and the poling process is complete. As long as the poled material is maintained below the Curie temperature, its polarization remains permanent. The dipoles stay "frozen" in the direction which was given to them by the electric field at high temperature (Fig. 3.23C).

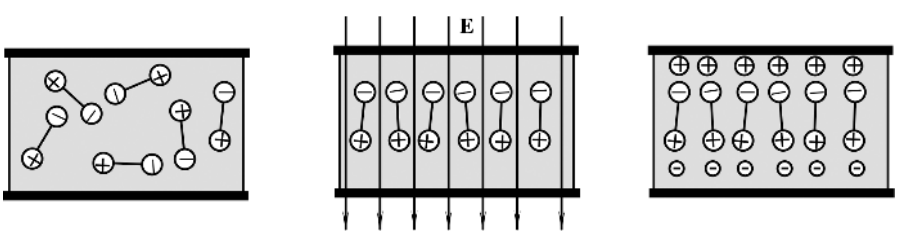


Fig. 3.23. Thermal poling of a piezoelectric and pyroelectric material.

Another method, called corona discharge poling, is also used to produce polymer piezo/pyroelectrics. The film is subjected to a corona discharge from an electrode at several million volts per centimeter of film thickness for 40–50 sec [5,6]. Corona polarization is uncomplicated to perform and can be easily applied before electric breakdown occurs, making this process useful at room temperature.

The final operation in preparation of the sensing element is shaping and finishing. This includes cutting, machining, and grinding. After the piezo (pyro) element is prepared, it is installed into a sensor's housing, where its electrodes are bonded to electrical terminals and other electronic components.

After poling, the crystal remains permanently polarized; however, it is electrically charged for a relatively short time. There is a sufficient amount of free carriers which move in the electric field setup inside the bulk material and there are plenty of charged ions in the surrounding air. The charge carriers move toward the poled dipoles and neutralize their charges (see Fig. 3.23C). Hence, after a while, the poled piezoelectric material becomes electrically discharged as long as it remains under steady-state conditions. When stress is applied, or air blows near its surface (Section 10.7 of Chapter 10) the balanced state is degraded and the piezoelectric material develops an electric charge. If the stress is maintained for a while, the charges again will be neutralized by the internal leakage. Thus, a piezoelectric sensor is responsive only to a changing stress rather than to a steady level of it. In other words, a piezoelectric sensor is an ac device, rather than a dc device.

Piezoelectric directional sensitivities (d coefficients) are temperature dependent. For some materials (quartz), the sensitivity drops with a slope of $-0.016\%/°\text{C}$. For others (the PVDF films and ceramics) at temperatures below 40°C, it may drop, and at higher temperatures, it increases with a raise in temperature. Currently, the most popular materials for fabrication of piezoelectric sensors are ceramics [7–9]. The earliest of the ferroelectric ceramics was barium titanate, a polycrystalline substance having the chemical formula $BaTiO_3$. The stability of permanent polarization relies on the coercive force of the dipoles. In some materials, polarization may decrease with time. To improve the stability of poled material, impurities have been introduced in the basic material with the idea that the polarization may be "locked" into position [4]. Although the piezoelectric constant changes with operating temperature, a dielectric constant, κ, exhibits a similar dependence. Thus, according to formula (3.71), variations in these values tend to cancel each other as they are entered into numerator and denominator. This results in a better stability of the output voltage, V, over a broad temperature range.

The piezoelectric elements may be used as a single crystal or in a multilayer form where several plates of the material are laminated together. This must be done with electrodes placed in between. Figure 3.24 shows a two-layer force sensor. When an external force is applied, the upper part of the sensor expands while the bottom compresses. If the layers are laminated correctly, this produces a double output signal. Double sensors can have either a parallel connection as shown in Fig. 3.25A or a serial connection as in Fig. 3.24C. The electrical equivalent circuit of the piezoelectric sensor is a parallel connection of a stress-induced current source (i), leakage resistance (r), and capacitance (C). Depending on the layer connection, equivalent circuits for the

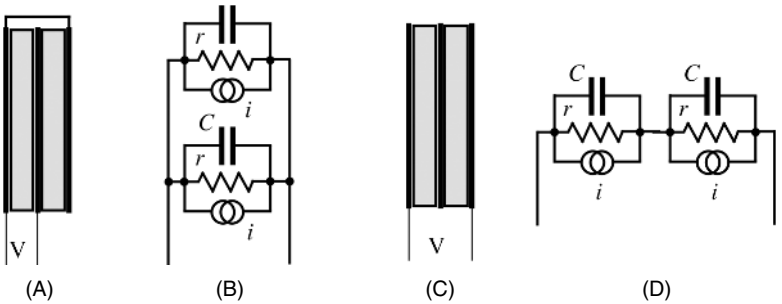


Fig. 3.24. Laminated two-layer piezoelectric sensor.

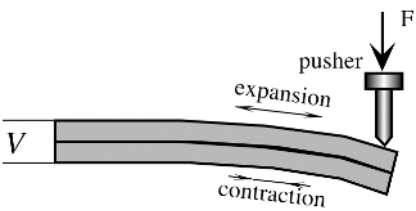
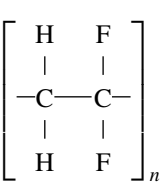


Fig. 3.25. Parallel (A) and serial (C) laminated piezoelectric sensors and their corresponding equivalent circuits (B and D).

laminated sensors are as shown in Figs. 3.25B and 3.25D. The leakage resistors r are very large (on the order of $10^{12} - 10^{14}\Omega$), which means that the sensor has an extremely high output impedance. This requires special interface circuits, such as charge and current-to-voltage converters, or voltage amplifiers with high input resistances.

3.6.1 Piezoelectric Films

In 1969, H. Kawai discovered a strong piezoelectricity in PVDF (polyvinylidene fluoride), and in 1975, the Japanese company Pioneer, Ltd. developed the first commercial product with the PVDF as piezoelectric loudspeakers and earphones [10]. PVDF is a semicrystalline polymer with an approximate degree of crystallinity of 50% [11]. Like other semicrystalline polymers, PVDF consists of a lamellar structure mixed with amorphous regions. The chemical structure of it contains the repeat unit of doubly fluorinated ethene CF_2–CH_2:



The molecular weight of PVDF is about 10^5, which corresponds to about 2000 repeat units. The film is quite transparent in the visible and near-IR (infrared) region and is

absorptive in the far-IR portion of the electromagnetic spectrum. The polymer melts at about 170°C. Its density is about 1780 kg/m^3. PVDF is a mechanically durable and flexible material. In piezoelectric applications, it is usually drawn, uniaxially or biaxially, to several times its length. Elastic constants, (e.g., Young's modulus) depend on this draw ratio. Thus, if the PVDF film was drawn at 140°C to the ratio of 4:1, the modulus value is 2.1 GPa, whereas for the draw ratio of 6.8:1, it was 4.1 GPa. The resistivity of the film also depends on the stretch ratio. For instance, at low stretch, it is about $6.3 \times 10^{15}\Omega$ cm, whereas for the stretch ratio 7:1 it is $2 \times 10^{16}\Omega$ cm.

Polyvinylidene fluoride does not have a higher or even as high piezoelectric coefficient as other commonly used materials, like $BaTiO_3$ or PZT. However, it has a unique quality not to depolarize while being subjected to very high alternating electric fields. This means that even though the value of d_{31} of PVDF is about 10% of PZT, the maximum strain observable in PVDF will be 10 times larger than in PZT because the maximum permissible field is 100 times greater for PVDF. The film exhibits good stability: When stored at 60°C, it loses its sensitivity by about 1–2% over 6 months.

Comparative characteristics for various piezoelectric materials are given in Table A.8. Another advantage of piezo film over piezo ceramic is its low acoustic impedance, which is closer to that of water, human tissue, and other organic materials. For example, the acoustic impedance of piezo film is only 2.6 times that of water, whereas piezo ceramics are typically 11 times greater. A close impedance match permits more efficient transduction of acoustic signals in water and tissue.

Some unique properties of the piezoelectric films are as follows[8]:

- Wide frequency range: 0.001 Hz to 10^9 Hz
- Vast dynamic range: 10^{-8}–10^6 psi or μtorr to Mbar.
- Low acoustic impedance: close match to water, human tissue, and adhesive systems
- High elastic compliance
- High voltage output: 10 times higher than piezo ceramics for the same force input
- High dielectric strength: withstanding strong fields (75 V/μm), where most piezo ceramics depolarize
- High mechanical strength and impact resistance: 10^9–10^{10} P modulus.
- High stability: resisting moisture (<0.02% moisture absorption), most chemicals, oxidants, and intense ultraviolet and nuclear radiation
- Can be fabricated into many shapes
- Can be glued with commercial adhesives

Typical properties of piezoelectric films are given in Table 3.1.

Like some other ferroelectric materials, PVDF is also pyroelectric (see Section 3.7), producing electrical charge in response to a change in temperature. PVDF strongly absorbs infrared energy in the 7–20-μm wavelength range, covering the same wavelength spectrum as heat from the human body. However, in spite of the fact that the film can absorb thermal radiation, a pyroelectric sensor has the film sandwiched between two thin metal electrodes, which can be quite reflective in the

[8] from Measurement Specialties, Inc. (www.msiusa.com).

Table 3.1. Typical Properties of Piezoelectric Films

Symbol	Parameter	PVDF	Copolymer	Units
t	Thickness	9, 28, 52, 110	<1 to 1200	μm (micron, 10^{-6})
d_{31}	Piezo Strain Constant	23	11	10^{-12} $\frac{m/m}{V/m}$ or $\frac{C/m^2}{N/m^2}$
d_{33}		−33	−38	
g_{31}	Piezo Stress Constant	216	162	10^{-3} $\frac{V/m}{N/m^2}$ or $\frac{m/m}{C/m^2}$
g_{33}		−330	−542	
k_{31}	Electromechanical Coupling Factor	12%	20%	
k_t		14%	25–29%	
C	Capacitance	380 for 28μm	68 for 100μm	pF/cm^2 1KHz
Y	Young's Modulus	2–4	3–5	10^9 N/m^2
V_0	Speed of Sound stretch:	1.5	2.3	10^3 m/s
	Speed of Sound thickness:	2.2	2.4	
p	Pyroelectric Coefficient	30	40	10^{-6} C/m^2 °K
ε	Permittivity	106–113	65–75	10^{-12} F/m
$\varepsilon/\varepsilon_0$	Relative Permittivity	12–13	7–8	
ρ_m	Mass Density	1.78	1.82	10^3kg/m
ρ_e	Volume Resistivity	$>10^{13}$	$>10^{14}$	Ohm meters
$R_\square$	Surface Metallization Resistivity	<3.0	<3.0	Ohms/square for NiAl
$R_\square$		0.1	0.1	Ohms/square for Ag Ink
$\tan\delta_e$	Loss Tangent	0.02	0.015	1KHz
	Yield Strength	45–55	20–30	10^6 N/m^2 (stretch axis)
	Temperature Range	−40 to 80 . . . 100	−40 to 115 . . . 145	°C
	Water Absorption	<0.02	<0.02	% H_2O
	Maximum Operating Voltage	750 (30)	750 (30)	V/mil(V/μm), DC, 25°C
	Breakdown Voltage	2000 (80)	2000 (80)	V/mil(V/μm), DC, 25°C

Source: Ref. [12].

spectral range of interest. In such cases, the electrode that is exposed to thermal radiation either is coated with a heat-absorbing layer or is made of nichrome (a metal having high absorptivity). PVDF makes a useful human motion sensor as well as pyroelectric sensor for more sophisticated applications like vidicon cameras for night vision and laser beam profiling sensors. A dense infrared array has been recently introduced that identifies one's fingerprint pattern using the pyro effect of the piezo polymer. New copolymers of PVDF, developed over the last few years, have expanded the applications of piezoelectric polymer sensors. These copolymers permit use at higher temperatures (135°C) and offer desirable new sensor shapes, like cylinders and hemispheres. Thickness extremes are possible with copolymer that cannot be readily attained with PVDF. These include ultrathin (200Å) spin-cast coatings

that enable new sensor-on-silicon applications, and cylinders with wall thicknesses in excess of 1200 μm for sonar. Piezo cable is also produced using a copolymer.

Unlike piezo ceramic transducers, piezo film transducers offer wide dynamic range and are also broadband. These wide-band characteristics (near dc to 2 GHz) and low Q are partly attributable to the polymers' softness. As audio transmitters, a curved piezo film element, clamped at each end, vibrates in the length (d_{31}) mode. The d_{31} configuration is also used for air ultrasound ranging applications up to frequencies of about 50 kHz. When used as a high ultrasonic transmitter (generally >500 kHz), piezo film is normally operated in the thickness (d_{33}) mode. Maximum transmission occurs at thickness resonance. The basic half-wavelength resonance of 28-μm piezo film is about 40 MHz: Resonance values depend on film thickness. They range from low megahertz for thick films (1000) to >100 MHz for very thin films (μm).

Piezo film does have some limitations for certain applications. It makes a relatively weak electromechanical transmitter when compared to ceramics, particularly at resonance and in low-frequency applications. The copolymer film has maximum operating/storage temperatures as high as 135°C, whereas PVDF is not recommended for use or storage above 100°C. Also, if the electrodes on the film are exposed, the sensor can be sensitive to electromagnetic radiation. Good shielding techniques are available for high-electromagnetic interferences/radio-frequency interferences (EMI/RFI) environments. Table 3.1 lists typical properties of piezo film. Table A.8 provides a comparison of the piezoelectric properties of PVDF polymer and other popular piezoelectric ceramic materials. Piezo film has low density and excellent sensitivity and is mechanically tough. The compliance of piezo film is 10 times greater than the compliance of ceramics. When extruded into thin film, piezoelectric polymers can be directly attached to a structure without disturbing its mechanical motion. Piezo film is well suited to strain-sensing applications requiring a very wide bandwidth and high sensitivity. As an actuator, the polymer's low acoustic impedance permits the efficient transfer of a broadband of energy into air and other gases.

The piezoelectric effect is the prime means of converting mechanical deformation into electrical signal and vice versa in the miniature semiconductor sensors. The effect, however, can be used only for converting the changing stimuli and cannot be used for conversion of steady-state or very slow-changing signals.

Because silicon does not possess piezoelectric properties, such properties can be added on by depositing crystalline layers of the piezoelectric materials. The three most popular materials are zinc oxide (ZnO), aluminum nitride (AlN), and the so-called solid-solution system of lead–zirconite–titanium oxides $Pb(Zr,Ti)O_3$ known as PZT ceramic, basically the same material used for fabrication of discrete piezoelectric sensors as described earlier.

Zinc oxide in addition to the piezoelectric properties also is pyroelectric. It was the first and most popular material for development of ultrasonic acoustic sensors, surface-acoustic-wave (SAW) devices, microbalances, and so forth. One of its advantages is the ease of chemical etching. The zinc oxide thin films are usually deposited on silicon by employing sputtering technology.

Aluminum nitride is an excellent piezoelectric material because of its high acoustic velocity and its endurance in humidity and high temperature. Its piezoelectric coeffi-

cient is somewhat lower than in ZnO but higher than in other thin-film piezoelectric materials, excluding ceramics. The high acoustic velocity makes it an attractive choice in the gigahertz frequency range. Usually, the AlN thin films are fabricated by using the chemical vapor deposition (CVD) or reactive molecular beam epitaxy (MBE) technologies. However, the drawback of using these deposition methods is the need for a high heating temperature (up to 1300°C) of the substrate.

The PZT thin films possesses a larger piezoelectric coefficient than ZnO or AlN and also a high pyroelectric coefficient, which makes it a good candidate for fabrication of the thermal radiation detectors. A great variety of deposition techniques is available for the PZT, among which are electron-beam evaporation [13], radio-frequency (RF) sputtering [14], ion-beam deposition [15], epitaxial growth by RF sputtering [16], magnetron sputtering [17], laser ablation [18], and sol-gel [19].

3.7 Pyroelectric Effect

Pyroelectric materials are crystalline substances capable of generating an electrical charge in response to heat flow. The pyroelectric effect is very closely related to the piezoelectric effect. Before going further, we recommend that the reader be familiar with Section 3.6.

Like piezoelectrics, the pyroelectric materials are used in the form of thin slices or films with electrodes deposited on the opposite sides to collect the thermally induced charges (Fig. 3.26). The pyroelectric sensor is essentially a capacitor which can be electrically charged by an influx of heat. The detector does not require any external electrical bias (excitation signal). It needs only an appropriate electronic interface

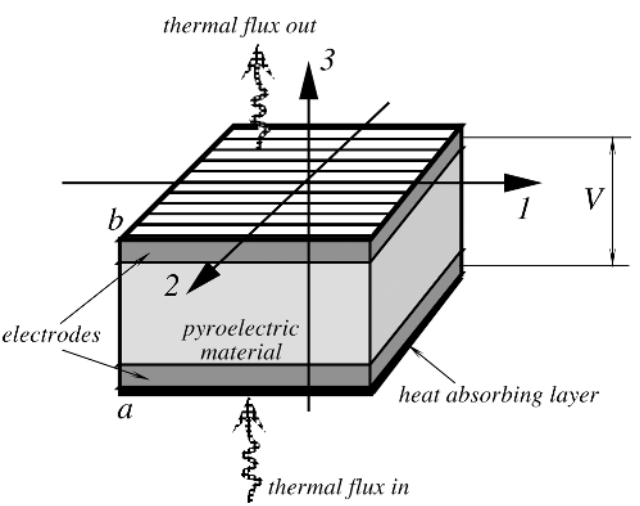


Fig. 3.26. The pyroelectric sensor has two electrodes on the opposite sides of the crystal. Thermal radiation is applied along axis 3 from the bottom and escapes upward.

circuit to measure the charge. Contrary to thermoelectrics (thermocouples) which produce a steady voltage when two dissimilar metal junctions are held at steady but different temperatures (see Section 3.9), pyroelectrics generate charge in response to a *change* in temperature. Because a change in temperature essentially requires propagation of heat, a pyroelectric device is a heat-*flow* detector rather than heat detector. Sometimes it is called a dynamic sensor, which reflects the nature of its response. When the pyroelectric crystal is exposed to a heat flow (e.g., from an infrared radiation source), its temperature elevates and it becomes a source of heat, in turn. Hence, there is an outflow of heat from the opposite side of the crystal, as is shown in Fig. 3.26.

A crystal is considered to be pyroelectric if it exhibits a spontaneous temperature-dependent polarization. Of the 32 crystal classes, 21 are noncentrosymmetric and 10 of these exhibit pyroelectric properties. In addition to pyroelectric properties, all of these materials exhibit some degree of piezoelectric properties as well: They generate an electrical charge in response to mechanical stress.

Pyroelectricity was observed for the first time in tourmaline crystals in the eighteenth century (some claim that the Greeks noticed it 23 centuries ago). Later, in the nineteenth century, Rochelle salt was used to make pyroelectric sensors. A large variety of materials became available after 1915: KDP (KH_2PO_4), ADP ($NH_4H_2PO_4$), $BaTiO_3$, and a composite of $PbTiO_3$ and $PbZrO_3$ known as PZT. Presently, more than 1000 materials with reversible polarization are known. They are called ferroelectric crystals. The most important among them are triglycine sulfate (TGS) and lithium tantalate ($LiTaO_3$). In 1969, H. Kawai discovered strong piezoelectricity in the plastic materials, polyvinyl fluoride (PVF) and polyvinylidene fluoride (PVDF) [20]. These materials also possess substantial pyroelectric properties.

A pyroelectric material can be considered as a composition of a large number of minute crystallities, each of which behaves as a small electric dipole. All of these dipoles are randomly oriented (Fig. 3.23A). Above a certain temperature, known as the *Curie point*, the crystallities have no dipole moment. Manufacturing (poling) of pyroelectric materials is analogous to that of piezoelectrics (see Section 3.6).

There are several mechanisms by which changes in temperature will result in pyroelectricity. Temperature changes may cause a shortening or elongation of individual dipoles. It may also affect the randomness of the dipole orientations due to thermal agitation. These phenomena are called *primary* pyroelectricity. There is also *secondary* pyroelectricity, which, in a simplified way, may be described as a result of the piezoelectric effect, (i.e., a development of strain in the material due to thermal expansion). Figure 3.26 shows a pyroelectric sensor whose temperature, T_0, is homogeneous over its volume. Being electrically polarized, the dipoles are oriented (poled) in such a manner as to make one side of the material positive and the opposite side negative. However, under steady-state conditions, free-charge carriers (electrons and holes) neutralize the polarized charge and the capacitance between the electrodes appears not to be charged (see Fig. 3.23C); that is, the sensor generates zero charge. Now, let us assume that heat is applied to the bottom side of the sensor. Heat may enter the sensor in a form of thermal radiation, which is absorbed by the bottom electrode and propagates toward the pyroelectric material via the mechanism of thermal

conduction. The bottom electrode may be given a heat-absorbing coating, such as goldblack or organic paint. As a result of heat absorption, the bottom side becomes warmer (the new temperature is T_1), which causes the bottom side of the material to expand. The expansion leads to flexing of the sensor, which, in turn, produces stress and a change in dipole orientation. Being piezoelectric, stressed material generates electric charges of opposite polarities across the electrodes. Hence, we may regard a secondary pyroelectricity as a sequence of events: a thermal radiation → a heat absorption → a thermally induced stress → an electric charge.

The dipole moment, M, of the bulk pyroelectric sensor is

$$M = \mu A h, \tag{3.72}$$

where μ is the dipole moment per unit volume, A is the sensor's area, and h is the thickness. The charge, Q_a, which can be picked up by the electrodes, develops the dipole moment across the material:

$$M_0 = Q_a h. \tag{3.73}$$

M must be equal to M_0, so that

$$Q_a = \mu A. \tag{3.74}$$

As the temperature varies, the dipole moment also changes, resulting in an induced charge.

Thermal absorption may be related to a dipole change, so that μ must be considered as a function of both temperature, T_a, and an incremental thermal energy, ΔW, absorbed by the material

$$\Delta Q_a = A\mu(T_a, \Delta W). \tag{3.75}$$

Figure 3.27 shows a pyroelectric detector (pyrosensor) connected to a resistor R_b, which represents either the internal leakage resistance or a combined input resistance of the interface circuit which is connected to the sensor. The equivalent electrical circuit of the sensor is shown at the right. It consists of three components: (1) the current source generating a heat induced current, i (remember that a current is a movement of electric charges), (2) the sensor's capacitance, C, and (3) the leakage resistance, R_b.

The output signal from the pyroelectric sensor can be taken in the form of either charge (current) or voltage, depending on the application. Being a capacitor, the pyroelectric device is discharged when connected to a resistor, R_b. Electric current

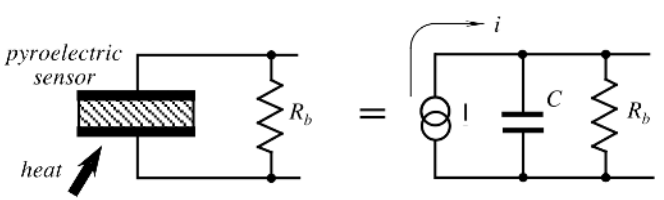


Fig. 3.27. Pyroelectric sensor and its equivalent circuit.

through the resistor and voltage across the resistor represent the heat-flow-induced charge. It can be characterized by two pyroelectric coefficients [21]:

$$\begin{aligned} P_Q &= \frac{dP_s}{dT} \quad \text{Pyroelectric charge coefficient,} \\ P_V &= \frac{dE}{dT} \quad \text{Pyroelectric voltage coefficient,} \end{aligned} \tag{3.76}$$

where P_s is the spontaneous polarization (which is another way to say *electric charge*), E is the electric field strength, and T is the temperature (in K). Both coefficients are related by way of the electric permitivity, ε_r, and dielectric constant, ε_0

$$\frac{P_Q}{P_V} = \frac{dP_s}{dE} = \varepsilon_r \varepsilon_0. \tag{3.77}$$

The polarization is temperature dependent and, as a result, both pyroelectric coefficients (3.76) are also functions of temperature.

If a pyroelectric material is exposed to a heat source, its temperature rises by ΔT and the corresponding charge and voltage changes can be described by the following equations:

$$\Delta Q = P_Q A \Delta T, \tag{3.78}$$

$$\Delta V = P_V h \Delta T. \tag{3.79}$$

Remembering that the sensor's capacitance can be defined as

$$C_e = \frac{\Delta Q}{\Delta V} = \varepsilon_r \varepsilon_0 \frac{A}{h}, \tag{3.80}$$

from (3.78–3.80) it follows that

$$\Delta V = P_Q \frac{A}{C_e} \Delta T = P_Q \frac{\varepsilon_r \cdot \varepsilon_0}{h} \Delta T. \tag{3.81}$$

It is seen that the peak output voltage is proportional to the sensor's temperature rise and pyroelectric charge coefficient and inversely proportional to its thickness.

When the pyroelectric sensor is subjected to a thermal gradient, its polarization (electric charge developed across the crystal) varies with the temperature of the crystal. A typical polarization–temperature curve is shown in Fig. 3.28. The voltage pyroelectric coefficient, P_v, is a slope of the polarization curve. It increases dramatically near the Curie temperature where the polarization disappears and the material permanently loses its pyroelectric properties. The curves imply that the sensor's sensitivity increases with temperature at the expense of nonlinearity.

To select the most appropriate pyroelectric material, the energy conversion efficiency should be considered. It is, indeed, the function of the pyroelectric sensor to convert thermal energy into electrical. "How effective is the sensor" is a key question in the design practice. A measure of efficiency is k_p^2, which is called the pyroelectric

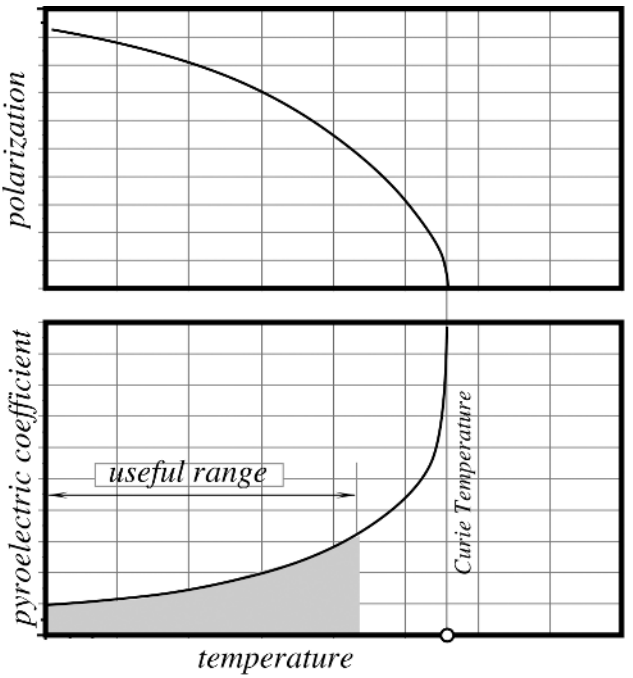


Fig. 3.28. Polarization of a pyroelectric crystal. The sensor must be stored and operated below the Curie temperature.

coupling coefficient[9] [21,22]. It shows the factor by which the pyroelectric efficiency is lower than the Carnot limiting value $\Delta T/T_a$. Numerical values for k_p^2 are shown in Table A.9.

Table A.9 represents that triglycine sulfate (TGS) crystals are the most efficient pyroelectric converters. However, for a long time, they were quite impractical for use in the sensors because of a low Curie temperature. If the sensor's temperature is elevated above that level, it permanently loses its polarization. In fact, TGS sensors proved to be unstable even below the Curie temperature, with a signal being lost quite spontaneously [23]. It was discovered that doping of TGS crystals with *L*-alanine (LATGS process patented by Philips) during its growth stabilizes the material below the Curie temperature. The Curie temperature was raised to 60°C, which allows its use with an upper operating temperature of 55°C, which is sufficient for many applications.

Other materials, such as lithium tantalate and pyroelectric ceramics, are also used to produce the pyroelectric sensors. Polymer films become increasingly popular for a variety of applications. During recent years, a deposition of pyroelectric thin films have been intensively researched. Especially promising is use of lead titanate ($PbTiO_3$), which is a ferroelectric ceramic having both a high pyroelectric coefficient

[9] The coefficient k_p is analogous to the piezoelectric coupling coefficient k.

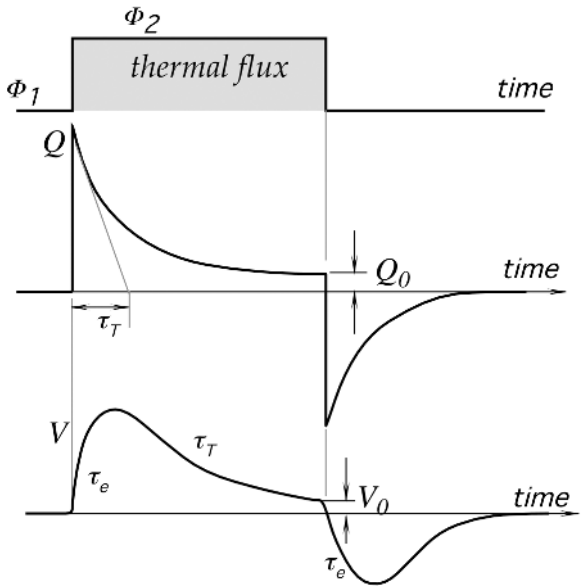


Fig. 3.29. Response of a pyroelectric sensor to a thermal step function. The magnitudes of charge Q_0 and voltage V_0 are exaggerated for clarity.

and a high Curie temperature of about 490°C. This material can be easily deposited on silicon substrates by the so-called sol-gel spin-casting deposition method [24].

Figure 3.29 shows the timing diagrams for a pyroelectric sensor when it is exposed to a step function of heat. It is seen that the electric charge reaches its peak value almost instantaneously, and then decays with a *thermal time constant*, τ_T. The physical meaning is this: A thermally induced polarization occurs initially in the most outer layer of the crystalline material (just few atomic layers), whose temperature nearly instantaneously raises to its maximum level. This creates the highest thermal gradient across the material thickness, leading to the maximum polarization. Then, heat starts propagating through the material, is being absorbed by its mass in proportion to its thermal capacitance C, and some of it is lost to the surroundings through thermal resistance R. This diminishes the initial gradient the generated charge. The thermal time constant is a product of the sensors' thermal capacitance and thermal resistance:

$$\tau_T = CR = cAhR, \tag{3.82}$$

where c is the specific heat of the sensing element. The thermal resistance R is a function of all thermal losses to the surroundings through convection, conduction, and thermal radiation. For the low-frequency applications, it is desirable to use sensors with τ_T as large as practical, whereas for high-speed applications (e.g., to measure laser pulses), a thermal time constant should be dramatically reduced. For that purpose, the pyroelectric material may be laminated with a heat sink (a piece of aluminum or copper).

When a pyroelectric sensor is exposed to a target, we consider a thermal capacity of a target very large (an infinite heat source) and the thermal capacity of the sensor small. Therefore, the surface temperature T_b of a target can be considered constant during the measurement, whereas the temperature of the sensor T_s is a function of time. That function is dependent on the sensing element: its density, specific heat, and thickness. If the input thermal flux has the shape of a step function of time, for the sensor freely mounted in air, the output current can be approximated by an exponential function, so that

$$i = i_0 e^{-t/\tau_T}, \tag{3.83}$$

where i_0 is peak current.

In Fig. 3.29, charge Q and voltage V do not completely return to zero, no matter how much time has elapsed. Thermal energy enters the pyroelectric material from side a (Fig. 3.26), resulting in a material temperature increase. This causes the sensor's response, which decays with a thermal time constant τ_T. However, because the other side, b, of the sensor faces a cooler environment, part of the thermal energy leaves the sensor and is lost to its surroundings. Because sides a and b face objects of different temperatures (one is the temperature of a target and the other is the temperature of the environment), a continuous heat flow exists through the pyroelectric material. The electric current generated by the pyroelectric sensor has the same shape as the thermal current through its material. An accurate measurement can demonstrate that as long as the heat continues to flow, the pyroelectric sensor will generate a constant voltage V_0 whose magnitude is proportional to the heat flow.

3.8 Hall Effect

This physical effect was discovered in 1879 at Johns Hopkins University by E. H. Hall. Initially, the effect had a limited, but very valuable application as a tool for studying electrical conduction in metals, semiconductors, and other conductive materials. Currently, Hall sensors are used to detect magnetic fields and position and displacement of objects [25,26].

The effect is based on the interaction between moving electric carriers and an external magnetic field. In metals, these carriers are electrons. When an electron moves through a magnetic field, a sideways force acts upon it:

$$\mathbf{F} = q v \mathbf{B}, \tag{3.84}$$

where $q = 1.6 \times 10^{-19}$ C is an electronic charge, v is the speed of an electron, and $\mathbf{B}$ is the magnetic field. Vector notations (boldface) are an indication that the force direction and its magnitude depend on the spatial relationship between the magnetic field and the direction of the electron movement. The unit of $\mathbf{B}$ is 1 tesla = 1 newton/(ampere meter) = 10^4 gauss.

Let us assume that the electrons move inside a flat conductive strip which is placed in a magnetic field $\mathbf{B}$ (Fig. 3.30). The strip has two additional contacts at its left and

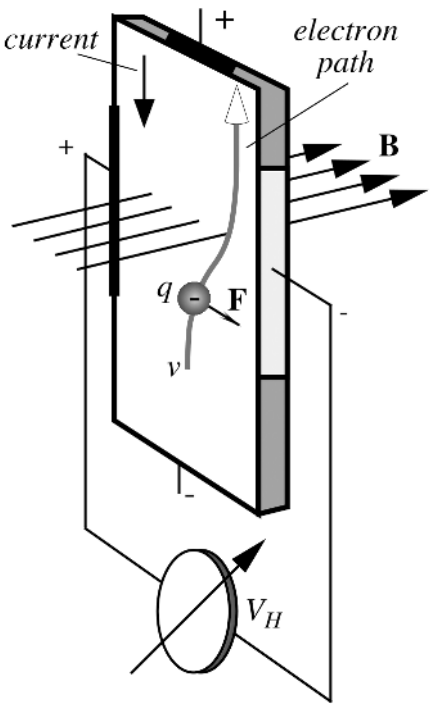


Fig. 3.30. Hall effect sensor. A magnetic field deflects movement of electric charges.

right sides which are connected to a voltmeter. Two other contacts are placed at the upper and lower ends of the strip. These are connected to a source of electric current. Due to the magnetic field, the deflecting force shifts moving electrons toward the right side of the strip, which becomes more negative than the left side; that is, the magnetic field and the electric current produce the so-called *transverse Hall potential difference* V_H. The sign and amplitude of this potential depends on both the magnitude and directions of magnetic field and electric current. At a fixed temperature, it is given by

$$V_H = hi\,B \sin\alpha, \tag{3.85}$$

where α is the angle between the magnetic field vector and the Hall plate (Fig. 3.31) and h is the coefficient of overall sensitivity whose value depends on the plate material, its geometry (active area), and its temperature.

The overall sensitivity depends on the *Hall coefficient*, which can be defined as the transverse electric potential gradient per unit magnetic field intensity per unit current density. According to the free-electron theory of metals, the Hall coefficient should be given by

$$H = \frac{1}{Ncq}, \tag{3.86}$$

where N is the number of free electrons per unit volume and c is the speed of light.

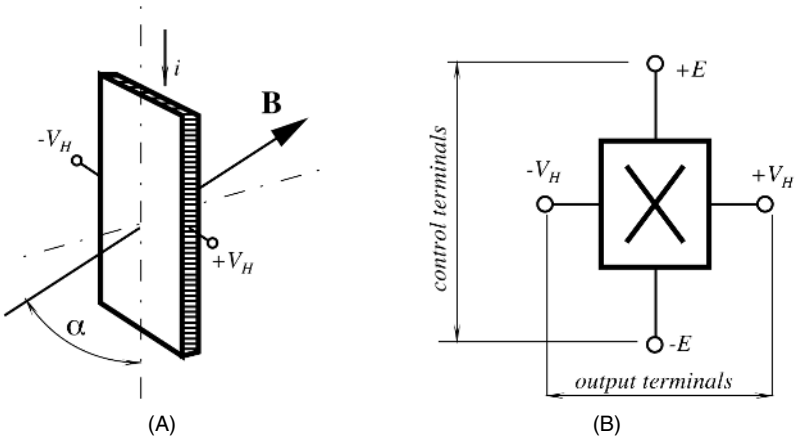


Fig. 3.31. The output signal of a Hall sensor depends on the angle between the magnetic field vector and the plate (A); four terminals of a Hall sensor (B).

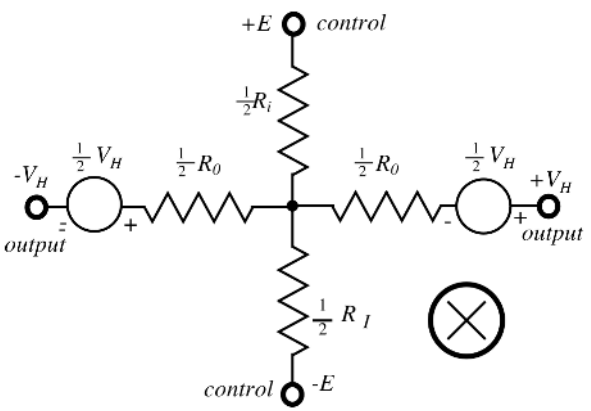


Fig. 3.32. Equivalent circuit of a Hall sensor.

Depending on the material crystalline structure, charges may be either electrons (negative) or holes (positive). As a result, the Hall effect may be either negative or positive.

A linear Hall effect sensor is usually packaged in a four-terminal housing. Terminals for applying the control current are called the *control* terminals and the resistance between them is called the *control resistance* R_i. Terminals where the output voltage is observed are called the *differential output terminals* and the resistance between them is called the *differential output resistance*, R_0. The sensor's equivalent circuit (Fig. 3.32) may be represented by cross-connected resistors and two voltage sources connected in series with the output terminals. The cross ⊗ in Figs. 3.31B and 3.32 indicates the direction of the magnetic field from the viewer to the symbol plane.

Table 3.2. Typical Characteristics of a Linear Hall Effect Sensor.

Control current	3 mA
Control resistance, R_i	2.2 kΩ
Control resistance versus temperature	+0.8%/°C
Differential output resistance, R_0	4.4 k Ω
Output offset voltage	5.0 mV (at $B = 0$ G)
Sensitivity	60 μV/G
Sensitivity versus temperature	+0.1%/°C
Overall sensitivity	20 V/ΩkG
Maximum magnetic flux density, B	Unlimited

Source: Ref. [27].

The sensor is specified by its resistances, R_i and R_0, across both pairs of terminals, the offset voltage at no magnetic field applied, the sensitivity, and the temperature coefficient of sensitivity. Many Hall effect sensors are fabricated from silicon and fall into two general categories: the basic sensors and the integrated sensors. Other materials used for the element fabrication include InSb, InAs, Ge, and GaAs. In the silicon element, an interface electronic circuit can be incorporated into the same wafer. This integration is especially important because the Hall effect voltage is quite low. For instance, a linear basic silicon sensor UGN-3605K manufactured by Sprague has typical characteristics presented in Table 3.2.

A built-in electronic interface circuit may contain a threshold device, thus making an integrated sensor a two-state device; that is, its output is "zero" when the magnetic field is below the threshold, and it is "one" when the magnetic field is strong enough to cross the threshold.

Because of the piezoresistivity of silicon, all Hall effect sensors are susceptible to mechanical stress effects. Caution should be exercised to minimize the application of stress to the leads or the housing. The sensor is also sensitive to temperature variations because temperature affects the resistance of the element. If the element is fed by a voltage source, the temperature will change the control resistance and, subsequently, the control current. Hence, it is preferable to connect the control terminals to a current source rather than to a voltage source.

One way to fabricate the Hall sensor is to use a silicon p-substrate with ion-implanted n-wells (Fig. 3.33A). Electrical contacts provide connections to the power-supply terminals and form the sensor outputs. A Hall element is a simple square with a well with four electrodes attached to the diagonals (Fig. 3.33B). A helpful way of looking at the Hall sensor is to picture it as a resistive bridge as depicted in Fig. 3.33C. This representation makes its practical applications more conventional because the bridge circuits are the most popular networks with well-established methods of design (Section 5.7 of Chapter 5).

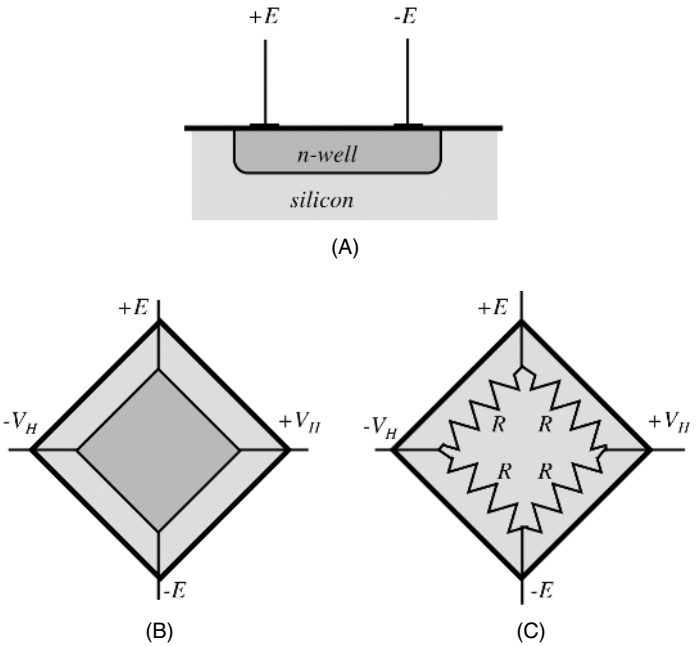


Fig. 3.33. Silicon Hall effect sensor with *n*-well (A and B) and its equivalent resistive bridge circuit (C).

3.9 Seebeck and Peltier Effects

In 1821, Thomas Johann Seebeck (1770–1831), an Estonian-born and Berlin- and Göttingen-educated physician, accidentally joined semicircular pieces of bismuth and copper while studying the thermal effects on galvanic arrangements [28]. A nearby compass indicated a magnetic disturbance (Fig. 3.34A). Seebeck experimented repeatedly with different metal combinations at various temperatures, noting related magnetic field strengths. Curiously, he did not believe that an electric current was flowing and preferred to describe that effect as "thermomagnetism" [29].

If we take a conductor and place one end of it into a cold place and the other end into a warm place, energy will flow from the warm to cold part. The energy takes the form of heat. The intensity of the heat flow is proportional to the thermal conductivity of the conductor. In addition, the thermal gradient sets an electric field inside the conductor (this directly relates to the Thompson effect[10]). The field results

[10] A Thompson effect was discovered by William Thompson around 1850. It consists of absorption or liberation of heat by passing current through a homogeneous conductor which has a temperature gradient across its length. The heat is linearly proportional to current. Heat is absorbed when current and heat flow in opposite directions, and heat is produced when they flow in the same direction.

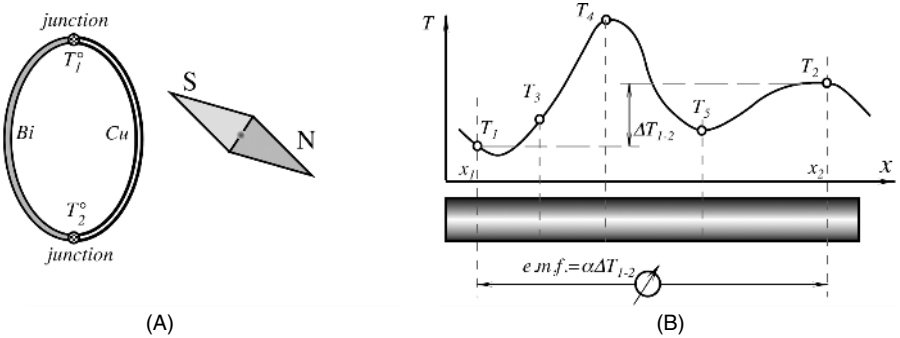


Fig. 3.34. (A) Seebeck experiment; (B) the varying temperature along a conductor is a source of a thermoelectric e.m.f.

in incremental voltage:

$$dV_a = \alpha_a \frac{dT}{dx}\, dx, \tag{3.87}$$

where dT is the temperature gradient across a small length, dx, and α_a is the *absolute* Seebeck coefficient of the material [30]. If the material is homogeneous, α_a is not a function of length and Eq. (3.87) reduces to

$$dV_a = \alpha_a\, dT. \tag{3.88}$$

Equation (3.88) is a principle mathematical expression of a thermoelectric effect. Figure 3.34B shows a conductor having nonuniform temperature T along its length x. A temperature gradient between any arbitrary points defines an electromotive force (e.m.f.) between these points. Other possible temperatures between the selected points (temperatures T_3, T_4 and T_5, for example) have no effect whatsoever on the value of e.m.f. between points 1 and 2. To measure the e.m.f., we connect a voltmeter to the conductor as shown in Fig. 3.34B; this is not as simple as may first look. To measure thermally induced e.m.f., we would need to attach the voltmeter probes. However, the probes are also made of conductors which may be different from the conductor we observe. Let us consider a simple measurement electric circuit where a current loop is formed. In such a loop, a meter is connected in series with the wire (Fig. 3.35A). If the loop is made of a uniform material, say cooper, then no current will be observed, even if the temperature along the conductor is not uniform. Electric fields in the left and right arms of the loop produce equal currents $i_a = i_b$, which cancel each other, resulting in a zero net current. A thermally induced e.m.f. exists in every thermally nonhomogeneous conductor, but it cannot be directly measured.

In order to observe *thermoelectricity*, it is, in fact, necessary to have a circuit composed of two *different* materials,[11] and we can then measure the *net* difference between their thermoelectric properties. Figure 3.35B shows a loop of two dissimilar metals which produces net current $\Delta i = i_a - i_b$. The actual current depends on many factors, including the shape and size of the conductors. If, on the other hand, instead

[11] Or perhaps the same material in two different states—for example, one under strain, and the other not.

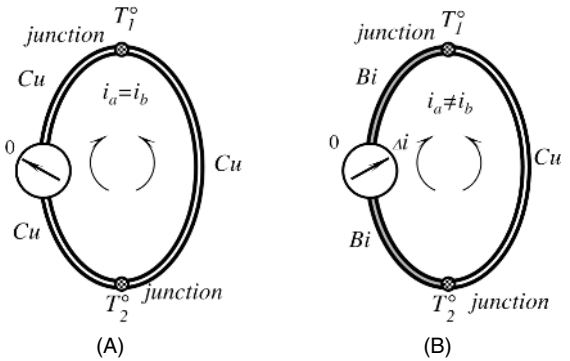


Fig. 3.35. Thermoelectric loop: (A) joints of identical metals produce zero net current at any temperature difference; (B) joints of dissimilar metals produce net current Δi.

of current we measure the net voltage across the broken conductor, the potential will depend *only* on the materials and the temperature difference. It does not depend on any other factors. A thermally induced potential difference is called the *Seebeck potential*.

What happens when two conductors are joined together? Free electrons in metal may behave as an ideal gas. The kinetic energy of electrons is a function of the material temperature. However, in different materials, energies and densities of free electrons are not the same. When two dissimilar materials at the same temperature are brought into a contact, free electrons diffuse through the junction [30]. The electric potential of the material accepting electrons becomes more negative at the interface, and the material emitting electrons becomes more positive. Different electronic concentrations across the junction set up an electric field which balances the diffusion process, and the equilibrium is established. If the loop is formed and both junctions are at the same temperature, the electric fields at both junctions cancel each other, which is not the case when the junctions are at different temperatures.

A subsequent investigation [40] has shown the Seebeck effect to be fundamentally electrical in nature. It can be stated that the thermoelectric properties of a conductor are, in general, just as much bulk properties as are the electrical and thermal conductivities. Coefficient α_a is a unique property of a material. When a combination of two dissimilar materials (A and B) is used, the Seebeck potential is determined from a *differential* Seebeck coefficient:

$$\alpha_{AB} = \alpha_A - \alpha_B, \tag{3.89}$$

and the net voltage of the junction is

$$dV_{AB} = \alpha_{AB}\, dT. \tag{3.90}$$

Equation (3.90) can be used to determine a differential coefficient:

$$\alpha_{AB} = \frac{dV_{AB}}{dT}. \tag{3.91}$$

For example, voltage as function of a temperature gradient for a T-type thermocouple with a high degree of accuracy can be approximated by a second-order equation

$$V_{AB} = a_0 + a_1 T + a_2 T^2 = -0.0543 + 4.094 \times 10^{-2} T + 2.874 \times 10^{-5} T^2; \quad (3.92)$$

then, a differential Seebeck coefficient for the T-type thermocouple is

$$\alpha_T = \frac{dV_{AB}}{dT} = a_1 + 2a_2 T = 4.094 \times 10^{-2} + 5.74810^{-5} T. \quad (3.93)$$

It is seen that the coefficient is a linear function of temperature. Sometimes, it is called the *sensitivity* of a thermocouple junction. A reference junction which is kept at a cooler temperature traditionally is called a *cold junction* and the warmer is a *hot junction*. The Seebeck coefficient does not depend on the nature of the junction: Metals may be pressed together, welded, fused, and so forth. What counts is the temperature of the junction and the actual metals. The Seebeck effect is a direct conversion of thermal energy into electric energy.

Table A.11 in the Appendix gives the values of thermoelectric coefficients and volume resistivities for some thermoelectric materials. It is seen that to achieve the best sensitivity, the junction materials should be selected with the opposite signs for α and those coefficients should be as large as practical.

In 1826, A. C. Becquerel suggested using Seebeck's discovery for temperature measurements. Nevertheless, the first practical thermocouple was constructed by Henry LeChatelier almost 60 years later [31]. He had found that the junction of platinum and platinum–rhodium alloy wires produce "the most useful voltage." Thermoelectric properties of many combinations have been well documented and for many years have been used for measuring temperature. Table A.10 (Appendix) gives the sensitivities of some practical thermocouples (at 25°C) and Fig. 3.36 shows the Seebeck voltages for the standard types of thermocouple over a broad temperature range. It should be emphasized that a thermoelectric sensitivity is not constant over the temperature range and it is customary to reference thermocouples at 0°C. In addition to the thermocouples, the Seebeck effect also is employed in *thermopiles*, which are, in essence, multiple serially connected thermocouples. Currently, thermopiles are most extensively used for the detection of thermal radiation (Section 14.6.2 of Chapter 14). The original thermopile was made of wires and was intended for increasing the output voltage. It was invented by James Joule (1818–1889) [32].

Currently, the Seebeck effect is used in the fabrication of integral sensors where pairs of materials are deposited on the surface of semiconductor wafers. An example is a thermopile which is a sensor for the detection of thermal radiation. Quite sensitive thermoelectric sensors can be fabricated of silicon, as silicon possess a strong Seebeck coefficient. The Seebeck effect results from the temperature dependence of the Fermi energy E_F, and the total Seebeck coefficient for n-type silicon may be approximated as a function of electrical resistivity for the range of interest (for use in sensors at room temperature):

$$\alpha_a = \frac{mk}{q} \ln\left(\frac{\rho}{\rho_0}\right), \quad (3.94)$$

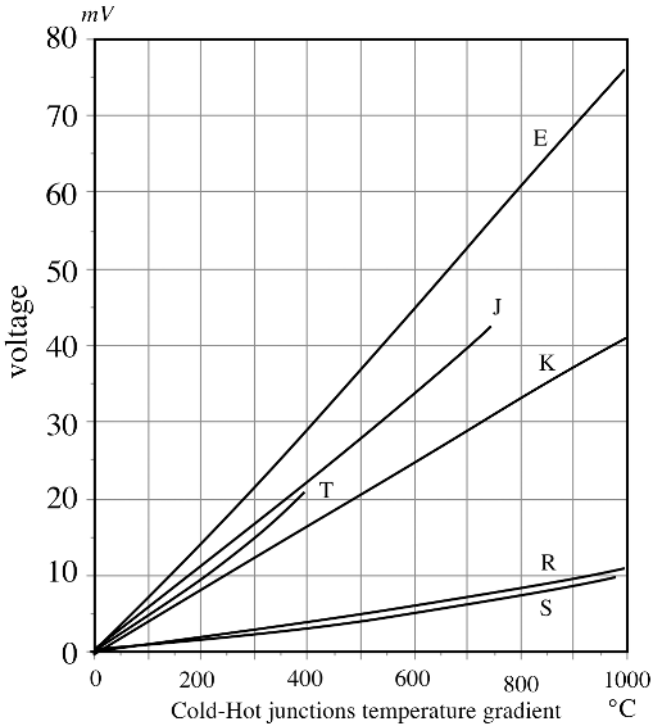


Fig. 3.36. Output voltage from standard thermocouples as functions of a cold–hot temperature gradient.

where $\rho_0 \approx 5 \times 10^{-6}\Omega$m and m ≈ 2.5 are constants, k is the Boltzmann constant, and q is the electronic charge. The doping concentrations used in practice lead to Seebeck coefficients on the order of 0.3–0.6 mV/K. The absolute Seebeck coefficients of a few selected metals and some typical values of silicon are shown in Table A.11. It can be seen that the Seebeck coefficients for metals are much smaller than for silicon and that the influence of aluminum terminals on chips is negligible compared to the Seebeck coefficient for silicon.

In the early nineteenth century, a French watchmaker turned physicist, Jean Charles Athanase Peltier (1785–1845), discovered that if electric current passes from one substance to another (Fig. 3.37), then heat may be given or absorbed at the junction [33]. Heat absorption or production is a function of the current direction:

$$dQ_P = \pm pi\, dt, \tag{3.95}$$

where i is the current and t is time. The coefficient p has a dimension of voltage and represents thermoelectric properties of the material. It should be noted that heat does not depend on the temperature at the other junction.

The Peltier effect concerns the reversible absorption of heat which usually takes place when an electric current crosses a junction between two dissimilar metals. The

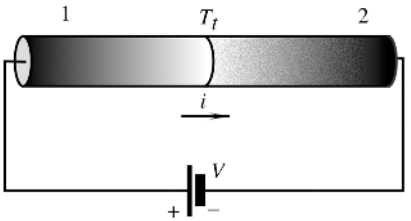


Fig. 3.37. Peltier effect.

effect takes place whether the current is introduced externally or is induced by the thermocouple junction itself (due to the Seebeck effect).

The Peltier effect is used for two purposes: It can produce heat or "produce" cold, depending on the direction of electric current through the junction. This makes it quite useful for the devices where precision thermal control is required. Apparently, the Peltier effect is of the same nature as the Seebeck effect. It should be well understood that the Peltier heat is different from that of the Joule. The Peltier heat depends *linearly* on the magnitude of the current flow as contrasted to Joule heat[12]. The magnitude and direction of Peltier heat do not depend in any way on the actual nature of the contact. It is purely a function of two different bulk materials which have been brought together to form the junction and each material makes its own contribution depending on its thermoelectric properties. The Peltier effect is a basis for operation of thermoelectric coolers, which are used for the cooling of photon detectors operating in the far-infrared spectral range (Section 14.5 of Chapter 14) and chilled mirror hygrometers (Section 13.5 of Chapter 13).

In summary, thermoelectric currents may exist whenever the junctions of a circuit formed of at least two dissimilar metals are exposed to different temperatures. This temperature difference is always accompanied by irreversible Fourier heat conduction, whereas the passage of electric currents is always accompanied by irreversible Joule heating effect. At the same time, the passage of electric current always is accompanied by reversible Peltier heating or cooling effects at the junctions of the dissimilar metals, whereas the combined temperature difference and passage of electric current always is accompanied by reversible Thomson heating or cooling effects along the conductors. The two reversible heating–cooling effects are manifestations of four distinct e.m.f.'s which make up the net Seebeck e.m.f.:

$$E_s = p_{AB|T_2} - p_{AB|T_1} + \int_{T_1}^{T_2} \sigma_A \, dT - \int_{T_1}^{T_2} \sigma_B \, dT = \int_{T_1}^{T_2} \alpha_{AB} \, dT, \qquad (3.96)$$

where σ is a quantity called the Thomson coefficient, which Thomson referred to as the specific heat of electricity, because of an apparent analogy between σ and the usual specific heat, c, of thermodynamics. The quantity of σ represents the rate at which heat is absorbed, or liberated, per unit temperature difference per unit mass [34,35].

[12] Joule heat is produced when electric current passes in any direction through a conductor having finite resistance. Released thermal power of Joule heat is proportional to squared current: $P = i^2/R$, where R is the resistance of a conductor.

3.10 Sound Waves

Alternate physical compression and expansion of medium (solids, liquids, and gases) with certain frequencies are called sound waves. The medium contents oscillate in the direction of wave propagation; hence, these waves are called longitudinal mechanical waves. The name *sound* is associated with the hearing range of a human ear, which is approximately from 20 to 20,000 Hz. Longitudinal mechanical waves below 20 Hz are called *infrasound* and above 20,000 Hz (20 kHz), they are called *ultrasound*. If the classification were made by other animals, like dogs, the range of sound waves surely would be wider.

Detection of infrasound is of interest with respect to analysis of building structures, earthquake prediction, and other geometrically large sources. When infrasound is of a relatively strong magnitude, it can be, if not heard, at least felt by humans, producing quite irritating psychological effects (panic, fear, etc.). Audible waves are produced by vibrating strings (string music instruments), vibrating air columns (wind music instruments), and vibrating plates (some percussion instruments, vocal cords, loudspeaker). Whenever sound is produced, air is alternatively compressed and rarefied. These disturbances propagate outwardly. A spectrum of waves may be quite different—from a simple monochromatic sounds from a metronome or an organ pipe, to a reach violin music. Noise may have a very broad spectrum. It may be of a uniform distribution of density or it may be "colored" with predominant harmonics at some of its portions.

When a medium is compressed, its volume changes from V to $V - \Delta V$. The ratio of change in pressure, Δp, to relative change in volume is called the bulk modulus of elasticity of medium:

$$B = -\frac{\Delta p}{\Delta V / V} = \rho_0 v^2, \tag{3.97}$$

where ρ_0 is the density outside the compression zone and v is the speed of sound in the medium. Then, the speed of sound can be defined as

$$v = \sqrt{\frac{B}{\rho_0}}. \tag{3.98}$$

Hence, the speed of sound depends on the elastic (B) and inertia (ρ_0) properties of the medium. Because both variables are functions of temperature, the speed of sound also depends on temperature. This feature forms a basis for operation of the acoustic thermometers (Section 16.5 of Chapter 16). For solids, longitudinal velocity can be defined through its Young's modulus E and Poisson ratio ν:

$$v = \sqrt{\frac{E(1-\nu)}{\rho_0(1+\nu)(1-2\nu)}}. \tag{3.99}$$

Table A.15 (Appendix) provides the speeds of longitudinal waves in some media. It should be noted that the speed depends on temperature, which always must be considered for the practical purposes.

If we consider the propagation of a sound wave in an organ tube, each small volume element of air oscillates about its equilibrium position. For a pure harmonic tone, the displacement of a particle from the equilibrium position may be represented by

$$y = y_m \cos \frac{2\pi}{\lambda}(x - vt), \tag{3.100}$$

where x is the equilibrium position of a particle and y is a displacement from the equilibrium position, y_m is the amplitude, and λ is the wavelength. In practice, it is more convenient to deal with pressure variations in sound waves rather than with displacements of the particles. It can be shown that the pressure exerted by the sound wave is

$$p = (k\rho_0 v^2 y_m) \sin(kx - \omega t), \tag{3.101}$$

where $k = 2\pi/\lambda$ is a wave number, ω is angular frequency, and the terms in the first parentheses represent an amplitude, p_m, of the sound pressure. Therefore, a sound wave may be considered a pressure wave. It should be noted that sin and cos in Eqs. (3.100) and (3.101) indicate that the displacement wave is 90° out of phase with the pressure wave.

Pressure at any given point in media is not constant and changes continuously, and the difference between the instantaneous and the average pressure is called the *acoustic pressure* P. During the wave propagation, vibrating particles oscillate near a stationary position with the instantaneous velocity ξ. The ratio of the acoustic pressure and the instantaneous velocity (do not confuse it with a wave velocity) is called the acoustic impedance:

$$Z = \frac{P}{\xi}, \tag{3.102}$$

which is a complex quantity, characterized by an amplitude and a phase. For an idealized media (no loss), $\mathbf{Z}$ is real and is related to the wave velocity as

$$Z = \rho_0 v. \tag{3.103}$$

We can define the intensity I of a sound wave as the power transferred per unit area. Also, it can be expressed through the acoustic impedance:

$$I = P\xi = \frac{P^2}{Z}. \tag{3.104}$$

It is common, however, to specify sound not by intensity but rather by a related parameter β, called the sound level and defined with respect to a reference intensity $I_0 = 10^{-12}$ W/m^2

$$\beta = 10 \log_{10}\left(\frac{I}{I_0}\right) \tag{3.105}$$

The magnitude of I_0 was chosen because it represents the lowest hearing ability of a human ear. The unit of β is a decibel (dB), named after Alexander Graham Bell. If $I = I_0$, $\beta = 0$.

Table 3.3. Sound Levels (β) Referenced to I_0 at 1000 Hz

Sound Source	dB
Rocket engine at 50 m	200
Supersonic boom	160
Hydraulic press at 1 m	130
Threshold of pain	120
10-W Hi-Fi speaker at 3 m	110
Unmuffled motorcycle	110
Rock-n-roll band	100
Subway train at 5 m	100
Pneumatic drill at 3 m	90
Niagara Falls	85
Heavy traffic	80
Automobiles at 5 m	75
Dishwashers	70
Conversation at 1 m	60
Accounting office	50
City street (no traffic)	30
Whisper at 1 m	20
Rustle of leaves	10
Threshold of hearing	0

Pressure levels also may be expressed in decibels as

$$\Pi = 20 \log_{10} \left(\frac{p}{p_0} \right), \tag{3.106}$$

where $p_0 = 2 \times 10^{-5}$ N/m^2 (0.0002 μbar)=2.9×10^{-9} psi.

Examples of some sound levels are given in Table 3.3. Because the response of a human ear is not the same at all frequencies, sound levels are usually referenced to I_0 at 1 kHz, at which the ear is most sensitive.

3.11 Temperature and Thermal Properties of Materials

Our bodies have a sense of temperature which by no means is an accurate method to measure outside heat. Human senses are not only nonlinear, but relative with respect to our previous experience. Nevertheless, we can easily tell the difference between warmer and cooler objects. Then, what is going on with these objects that they produce different perceptions?

Every single particle in this universe exists in perpetual motion. Temperature, in the simplest way, can be described as a measure of kinetic energy of vibrating particles. The stronger the movement, the higher the temperature of that particle. Of course, molecules and atoms in a given volume of material do not move with

equal intensities; that is, microscopically, they all are at different temperatures. The average kinetic energy of a large number of moving particles determines the *macroscopic* temperature of an object. These processes are studied by statistical mechanics. Here, however, we are concerned with methods and devices capable of measuring the macroscopic average kinetic energy of material particles, which is another way to state the temperature of the material. Because temperature is related to the movement of molecules, it is closely associated with pressure, which is defined as the force applied by moving molecules per unit area.

When atoms and molecules in a material move, they interact with other materials which happen to be brought in contact with them. Furthermore, every vibrating atom acts as a microscopic radiotransmitter which emanates electromagnetic radiation to the surrounding space. These two types of activity form a basis for heat transfer from warmer to cooler objects. The stronger the atomic movement, the hotter the temperature and the stronger the electromagnetic radiation. A special device (we call it a *thermometer*) which either contacts the object or receives its electromagnetic radiation produces a physical reaction, or signal. That signal becomes a measure of the object's temperature.

The word *thermometer* first appeared in literature in 1624 in a book by J. Leurechon, entitled *La Récréation Mathématique* [30]. The author described a glass water-filled thermometer whose scale was divided by 8 degrees. The first pressure-independent thermometer was built in 1654 by Ferdinand II, Grand Duke of Tuscany in a form of an alcohol-filled hermetically sealed tube.

Thermal energy is what we call heat. Heat is measured in *calories*[13]. One calorie (cal) is equal to the amount of heat which is required to warm up, by 1°C, 1 g of water at normal atmospheric pressure. In the United States, a British unit of heat is generally used, which is 1 Btu (British thermal unit): 1 Btu = 252.02 cal.

3.11.1 Temperature Scales

There are several scales for measuring temperature. The first zero for a scale was established in 1664 by Robert Hooke at a point of freezing distilled water. In 1694, Carlo Renaldi of Padua suggested taking the melting point of ice and the boiling point of water to establish two fixed points on a linear thermometer scale. He divided the span into 12 equal parts. Unfortunately, his suggestion had been forgotten for almost 50 years. In 1701, Newton also suggested to use two fixed points to define a temperature scale. For one point, he selected the temperature of melting ice (the zero point), and for the second point, he chose the armpit temperature of a healthy Englishman (he labeled that point 12). At Newton's scale, water was boiling at point No. 34. Daniel Gabriel Fahrenheit, a Dutch instrument maker, in 1706 selected *zero* for his thermometer at the coldest temperature produced by a mixture of water, ice, and sal-ammoniac or household salt. For the sake of a finer division, he established

[13] A *calorie* which measures energy in food is actually equal to 1000 physical calories, which is called a *kilocalorie*.

the other point at 96 degrees which is "…found in the blood of a healthy man…".[14] On his scale, the melting point of water was at 32° and boiling at 212°. In 1742, Andreas Celsius, professor of astronomy at the University of Uppsala, proposed a scale with zero as the melting point of ice and 100 at the boiling point of water.

Currently, in science and engineering, Celsius and Kelvin scales are generally employed. The Kelvin scale is arbitrarily based on the so-called *triple point of water*. There is a fixed temperature at a unique pressure of 4.58 mm Hg, where water vapor, liquid, and ice can coexist. This unique temperature is 273.16 K (kelvin) which approximately coincides with 0°C. The Kelvin scale is linear with zero intercept (0 K) at a lowest temperature where the kinetic energy of all moving particles is equal to zero. This point cannot be achieved in practice and is a strictly theoretical value. It is called the *absolute zero*. Kelvin and Celsius scales have the same slopes[15] (i.e., 1°C=1 K and 0 K=−273.15°C):

$$^{\circ}\mathrm{C} = ^{\circ}\mathrm{K} - 273.15^{\circ} \tag{3.107}$$

The boiling point of water is at 100°C=373.15 K. A slope of the Fahrenheit scale is steeper, because 1°C=1.8°F. The Celsius and Fahrenheit scales cross at temperature of −40°C and °F. The conversion between the two scales is

$$^{\circ}\mathrm{F} = 32 + 1.8^{\circ}\mathrm{C}, \tag{3.108}$$

which means that at 0°C, temperature in the Fahrenheit scale is +32°F.

3.11.2 Thermal Expansion

Essentially, all solids expand in volume with an increase in temperature. This is a result of vibrating atoms and molecules. When the temperature goes up, an average distance between the atoms increases, which leads to an expansion of a whole solid body. The change in any linear dimension (length, width, or height) is called a *linear expansion*. A length, l_2, at temperature, T_2, depends on length, l_1, at initial temperature T_1:

$$l_2 = l_1[1 + \alpha(T_2 - T_1)], \tag{3.109}$$

where α, called the coefficient of linear expansion, has different values for different materials. It is defined as

$$\alpha = \frac{\Delta l}{l}\frac{1}{\Delta T}, \tag{3.110}$$

[14] After all, Fahrenheit was a toolmaker and for him 96 was a convenient number because to engrave the graduation marks he could easily do so by dividing by 2: 96, 48, 24, 12, and so forth. With respect to nationality of the blood, he did not care if it was of Englishman or not. Now, it is known that the blood temperature of a healthy person is not really constant and varies between approximately 97°F and 100°F (36°C and 37.7°C), but during his time he did not have a better thermostat than the human body.

[15] There is a difference of 0.01° between Kelvin and Celsius scales, as Celsius' zero point is defined not at a triple point of water as for the Kelvin, but at temperature where ice and air-saturated water are at equilibrium at atmospheric pressure.

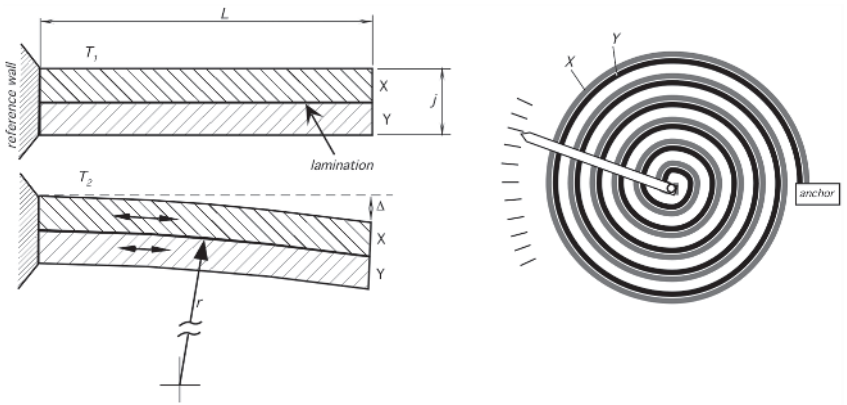


Fig. 3.38. (A) Warping of a laminated plate where two materials have different coefficients of thermal expansion; (B) a bimetal coil used as a temperature transducer.

where $\Delta T = T_2 - T_1$. Table A.16 gives values of α for different materials[16]. Strictly speaking, α depends on the actual temperature. However, for most engineering purposes, small variations in α may be neglected. For the so-called *isotropic* materials, α is the same for any direction. The fractional change in area of an object and its volume with a high degree of accuracy can be represented, respectively, by

$$\Delta A = 2\alpha A \Delta T, \tag{3.111}$$

$$\Delta V = 3\alpha V \Delta T. \tag{3.112}$$

Thermal expansion is a useful phenomenon that can be employed in many sensors where thermal energy is either measured or used as an excitation signal. Consider two laminated plates, X and Y, that are fused together (Fig. 3.38A). The plates have the same thickness and surface area and identical moduli of elasticity. Their coefficients of thermal expansion, α_1 and α_2, however, are different. The fused plates are anchored at the left-hand side to the reference wall. Now, if we apply heat to the structure, (i.e., if we increase its temperature from T_1 to T_2), plate X will expand more than plate Y (for $\alpha_1 > \alpha_2$). The lamination area will restrain plate X from a uniform expansion while forcing plate Y to expand more than its coefficient of expansion would require. This results in the formation of the internal stress and the structure will warp downward. Contrary, if we cool the structure, it will warp upward. The radius of warping can be estimated from [36]

$$r \approx \frac{2j}{3(\alpha_X - \alpha_Y)(T_2 - T_1)}. \tag{3.113}$$

The warping results in deflection of the tip that is strongest at the end of the structure anchored at the other end. This deflection can be measured as a representative of the

[16] More precisely, thermal expansion can be modeled by higher-order polynomials: $l_2 = l_1[1 + \alpha_1(T_2 - T_1) + \alpha_2(T_2 - T_1)^2 + \alpha_3(T_2 - T_1)^3 + \cdots]$; however, for the majority of practical purposes, a linear approximation is usually sufficient.

temperature change. This assumes that at a reference temperature (we may call it calibration temperature), the plate is flat; however, any convenient shape at a calibration temperature may be selected. In effect, a bimetal plate is a transducer of temperature into a displacement.

Most of such transducers are made of the bimetal plates (iron–nickel–chrome alloys). They are useful in a temperature range from −75°C and up to +600°C. In reality, for relatively small temperature changes, the radius of curvature is quite large (several meters) and thus the tip deflection is rather small. A bimaterial plate deflection can be computed from

$$\Delta = r\left[1 - \cos\left(\frac{180L}{\pi r}\right)\right], \tag{3.114}$$

where r is found from Eq. (3.113) and L is the length of the plate. For example, for a bimetal plate of $L = 50$ mm and thickness $j = 1$ mm and made of brass ($\alpha = 20 \times 10^{-6}$) and chromium ($\alpha = 6 \times 10^{-6}$) with a 10°C gradient, the deflection $\Delta \approx 0.26$ mm. This deflection is not easy to observe with the naked eye; thus, in a practical thermometer, a bimetal plate is usually preshaped in the form of a coil (Fig. 3.38B). This allows for a dramatic increase in L and achieve a much larger Δ. In the same example, for $L = 200$ mm, the deflection becomes 4.2 mm—a significant improvement. In modern sensors, the bimaterial structure is fabricated by employing a micromachining technology.

3.11.3 Heat Capacity

When an object is warmed, its temperature increases. By warming, we mean the transfer of a certain amount of heat (thermal energy) into the object. Heat is stored in the object in the form of the kinetic energy of vibration atoms. The amount of heat which an object can store is analogous to the amount of water which a water tank can store. Naturally, it cannot store more than its volume, which is a measure of a tank's capacity. Similarly, every object may be characterized by a heat capacity which depends on both the material of the object and its mass, m:

$$C = cm, \tag{3.115}$$

where c is a constant which characterizes the thermal properties of material. It is called the *specific heat* and is defined as

$$c = \frac{Q}{m\Delta T} \tag{3.116}$$

The specific heat describes the material, whereas a thermal capacity describes an object made of that material. Strictly speaking, specific heat is not constant over an entire temperature range of the specific phase of the material. It may change dramatically when a phase of the material changes, say from solid to liquid. Microscopically, specific heat reflects structural changes in the material. For instance, the specific heat of water is almost constant between 0°C and 100°C (liquid phase)—almost, but not exactly: It is higher near freezing and decreases slightly when the temperature goes

to about 35°C and then slowly rises again from 38°C to 100°C. Remarkably, the specific heat of water is the lowest near 37°C: the biologically optimal temperature of the warm-blooded animals.

Table A.17 gives the specific heats for various materials in cal/(g °C). Other tables provide specific heat in SI units of energy, which is, J/g °C. The relationship between cal/(g °C) and J/(g °C) is as follows

$$1\frac{\text{j}}{\text{g°C}} = 0.2388\frac{\text{cal}}{\text{g°C}}. \tag{3.117}$$

It may be noted that, generally, the heavier the material, the lower is its specific heat.

3.12 Heat Transfer

There are two fundamental properties of heat which should be well recognized:

(1) The heat is totally *not specific*; that is, once it is produced, it is impossible to say what origin it has.

(2) The heat *cannot be contained*, which means that it flows spontaneously from the warmer part to the cooler part of the system.

Thermal energy may be transferred from one object to another in three ways: conduction, convection, and radiation. Naturally, one of the objects which gives or receives heat may be a thermal detector. Its purpose would be to measure the amount of heat which represents some information about the object producing that heat. Such information may be the temperature of an object, chemical reaction, location or movement of the object, and so forth.

Let us consider a sandwichlike multilayer entity, where each layer is made of a different material. When heat moves through the layers, a temperature profile within each material depends on its thickness and thermal conductivity. Figure 3.39 shows three laminated layers where the first layer is attached to a heat source (a device having an "infinite" heat capacity and a high thermal conductivity). One of the best solid materials to act as an infinite heat source is a thermostatically controlled bulk copper. The temperature within the source is higher and constant, except of a very thin region near the laminated materials. Heat propagates from one material to another by conduction. The temperature within each material drops with different rates depending on the thermal properties of the material. The last layer loses heat to air through natural convection and to the surrounding objects through infrared radiation. Thus, Fig. 3.39 illustrates all three possible ways to transfer heat from one object to another.

3.12.1 Thermal Conduction

Heat conduction requires a physical contact between two bodies. Thermally agitated particles in a warmer body jiggle and transfer kinetic energy to a cooler body by agitating its particles. As a result, the warmer body loses heat while the cooler body gains heat. Heat transfer by conduction is analogous to water flow or to electric current.

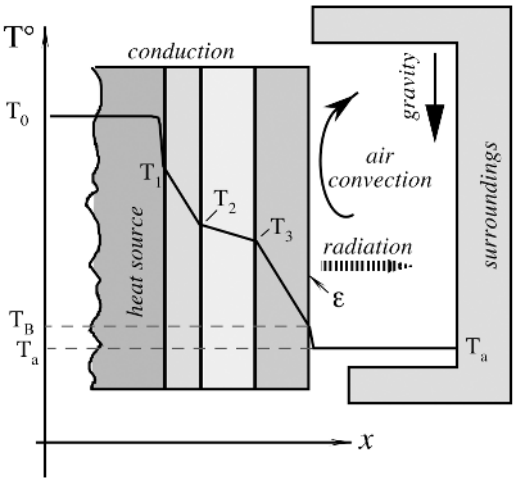


Fig. 3.39. Temperature profile in laminated materials.

For instance, heat passage through a rod is governed by a law which is similar to Ohm's law. The heat flow rate (thermal "current") is proportional to the thermal gradient (thermal "voltage") across the material (dT/dx) and the cross-sectional area A:

$$H = \frac{dQ}{dt} = -kA\frac{dT}{dx}, \tag{3.118}$$

where k is called *thermal conductivity*. The minus sign indicates that heat flows in the direction of temperature decrease (a negative derivative is required to cancel the minus sign). A good thermal conductor has a high k (most of metals), whereas thermal insulators (most of dielectrics) have a low k. Thermal conductivity is considered constant; however, it increases somewhat with temperature. To calculate a heat conduction through, say, an electric wire, temperatures at both ends (T_1 and T_2) must be used in equation

$$H = kA\frac{T_1 - T_2}{L}, \tag{3.119}$$

where L is the length of the wire. Quite often, thermal resistance is used instead of thermal conductivity:

$$R = \frac{L}{k}, \tag{3.120}$$

Then, Eq. (3.119) can be rewritten

$$H = A\frac{T_1 - T_2}{R}. \tag{3.121}$$

Values of thermal conductivities for some materials are shown in Table A.17.

Figure 3.40 shows an idealized temperature profile within the layers of laminated materials having different thermal conductivities. In the real world, heat transfer

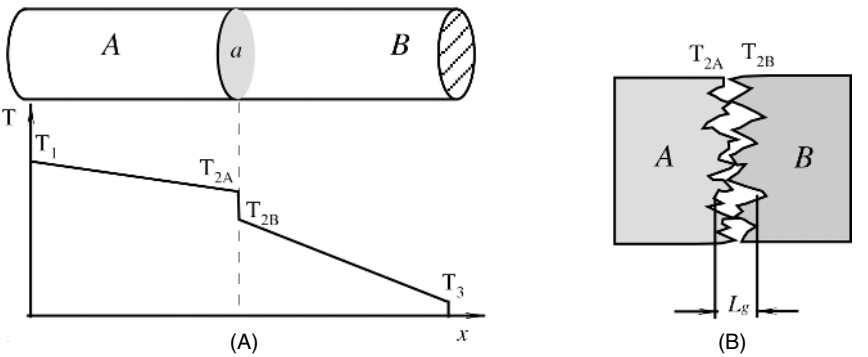


Fig. 3.40. Temperature profile in a joint (A) and a microscopic view of a surface contact (B).

through an interface of two adjacent materials may be different from the idealized case. If we join together two materials and observe the heat propagation through the assembly, a temperature profile may look like the one shown in Fig. 3.40A. If the sides of the materials are well insulated, under steady-state conditions, the heat flux must be the same through both materials. The sudden temperature drop at the interface, having surface area, a, is the result of a thermal *contact resistance*. Heat transfer through the assembly can be described as

$$H = \frac{T_1 - T_3}{R_A + R_c + R_B}, \tag{3.122}$$

where R_A and R_B are thermal resistances of two materials and R_c is the contact resistance,

$$R_c = \frac{1}{h_c a}. \tag{3.123}$$

The quantity h_c is called the contact coefficient. This factor can be very important in a number of sensor applications because many heat-transfer situations involve the mechanical joining of two materials. Microscopically, the joint may look like the one shown in Fig. 3.40B. No real surface is perfectly smooth, and the actual surface roughness is believed to play a central role in determining the contact resistance. There are two principal contributions to the heat transfer at the joint:

1. The material-to-material conduction through the actual physical contact
2. The conduction through trapped gases (air) in the void spaces created by the rough surfaces

Because the thermal conductivity of gases is very small compared with many solids, the trapped gas creates the most resistance to heat transfer. Then, the contact coefficient can be defined as

$$h_c = \frac{1}{L_g}\left(\frac{a_c}{a}\frac{2k_A k_B}{k_A + k_B} + \frac{a_v}{a}k_f\right), \tag{3.124}$$

where L_g is the thickness of the void space, k_f is the thermal conductivity of the fluid (e.g., air) filling the void space, a_c and a_v are areas of the contact and void, respectively,

and k_A and k_B are the respective thermal conductivities of the materials. The main problem with this theory is that it is very difficult to determine experimentally areas a_c and a_v and distance L_g. This analysis, however, allows us to conclude that the contact resistance should increase with a decrease in the ambient gas pressure. On the other hand, contact resistance decreases with an increase in the joint pressure. This is a result of a deformation of the high spots of the contact surface, which leads to enlarging a_c and creating a greater contact area between the materials. To decrease the thermal resistance, a dry contact between materials should be avoided. Before joining, surfaces may be coated with fluid having low thermal resistance. For instance, silicone thermal grease is often used for the purpose.

3.12.2 Thermal Convection

Another way to transfer heat is convection. It requires an intermediate agent (fluid: gas or liquid) that takes heat from a warmer body, carries it to a cooler body, releases heat, and then may or may not return back to a warmer body to pick up another portion of heat. Heat transfer from a solid body to a moving agent or within the moving agent is also called convection. Convection may be natural (gravitational) or forced (produced by a mechanism). With the natural convection of air, buoyant forces produced by gravitation act upon air molecules. Warmed-up air rises, carrying heat away from a warm surface. Cooler air descends toward the warmer object. Forced convection of air is produced by a fan or blower. Forced convection is used in liquid thermostats to maintain the temperature of a device at a predetermined level. The efficiency of a convective heat transfer depends on the rate of media movement, temperature gradient, surface area of an object, and thermal properties of moving medium. An object whose temperature is different from the surroundings will lose (or receive) heat, which can be determined from an equation similar to that of thermal conduction:

$$H = \alpha A(T_1 - T_2), \tag{3.125}$$

where convective coefficient α depends on the fluid's specific heat, viscosity, and a rate of movement. The coefficient is not only gravity dependent, but its value changes somewhat with the temperature gradient. For a horizontal plate in air, the value of α may be estimated from

$$\alpha = 2.49\sqrt[4]{T_1 - T_2}\,\frac{\mathrm{W}}{\mathrm{m^2K}}, \tag{3.126}$$

whereas for a vertical plate, it is

$$\alpha = 1.77\sqrt[4]{T_1 - T_2}\,\frac{\mathrm{W}}{\mathrm{m^2K}}. \tag{3.127}$$

It should be noted, however, that these values are applicable for one side of a plate only, assuming that the plate is a surface of an infinite heat source (i.e., its temperature does not depend on heat loss) and the surroundings have constant temperature. If the volume of air is small, like in the air gap between two surfaces of different temperatures, movement of gaseous molecules becomes very restricted and convective heat transfer

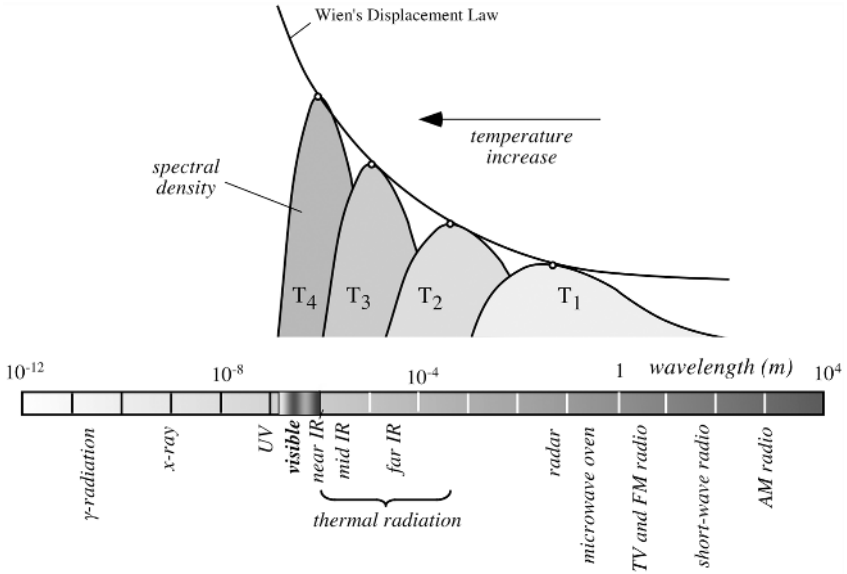


Fig. 3.41. Spectrum of electromagnetic radiation.

becomes insignificant. In these cases, thermal conductivity of air and radiative heat transfer should be considered instead.

3.12.3 Thermal Radiation

It was mentioned earlier that in any object, every atom and every molecule vibrate. The average kinetic energy of vibrating particles is represented by the absolute temperature. According to laws of electrodynamics, a moving electric charge is associated with a variable electric field that produces an alternating magnetic field. In turn, when the magnetic field changes, it results in a changing electric field coupled with it and so on. Thus, a vibrating particle is a source of an electromagnetic field which propagates outwardly with the speed of light and is governed by the laws of optics. Electromagnetic waves can be reflected, filtered, focused, and so forth. Figure 3.41 shows the total electromagnetic radiation spectrum which spreads from γ-rays to radio waves.

The wavelength directly relates to frequency, ν, by means of the speed of light c in a particular media:

$$\lambda = \frac{c}{\nu}. \tag{3.128}$$

A relationship between λ and temperature is more complex and is governed by Planck's law, which was discovered in 1901.[17] It establishes radiant flux density

[17] In 1918, Max K. E. L. Planck (Germany, Berlin University) was awarded the Nobel Prize "in recognition of his services he rendered to the advancement of Physics by his discovery of energy quanta."

W_λ as a function of wavelength λ and absolute temperature T. Radiant flux density is the power of electromagnetic radiation per unit of wavelength:

$$W_\lambda = \frac{\varepsilon(\lambda) C_1}{\pi \lambda^5 (e^{C_2/\lambda T} - 1)}, \tag{3.129}$$

where $\varepsilon(\lambda)$ is the emissivity of an object, $C_1 = 3.74 \times 10^{-12}$ Wcm^2 and $C_2 = 1.44$ cmK are constants, and e is the base of natural logarithms.

Temperature is a result of averaged kinetic energies of an extremely large number of vibrating particles. However, all particles do not vibrate with the same frequency or magnitude. Different permissive frequencies (also wavelengths and energies) are spaced very close to one another, which makes the material capable of radiating in a virtually infinite number of frequencies spreading from very long to very short wavelengths. Because temperature is a statistical representation of an average kinetic energy, it determines the highest probability for the particles to vibrate with a specific frequency and to have a specific wavelength. This most probable wavelength is established by Wien's law,[18] which can be found by equating to zero the first derivative of Eq. (3.129). The result of the calculation is a wavelength near which most of the radiant power is concentrated:

$$\lambda_m = \frac{2898}{T}, \tag{3.130}$$

where λ_m is in μm and T is in K. Wien's law states that the higher the temperature, the shorter the wavelength (Fig. 3.41). In view of Eq. (3.128), the law also states that the most probable frequency in the entire spectrum is proportional to the absolute temperature:

$$\nu_m = 10^{11} T \text{Hz}. \tag{3.131}$$

For instance, at normal room temperature, most of the far infrared energy is radiated from objects near 30 THz (30×10^{12} Hz). Radiated frequencies and wavelengths depend only on temperature, whereas the magnitude of radiation also depends on the emissivity $\varepsilon(\lambda)$ of the surface.

Theoretically, a thermal radiation bandwidth is infinitely wide. However, when detecting that radiation, properties of the real-world sensors must be taken into account. The sensors are capable of measuring only a limited range of radiation. In order to determine the total radiated power within a particular bandwidth, Eq. (3.129) is integrated within the limits from λ_1 to λ_2:

$$\Phi_{b0} = \frac{1}{\pi} \int_{\lambda_1}^{\lambda_2} \frac{\varepsilon(\lambda) C_1 \lambda^{-5}}{e^{C_2/\lambda T} - 1}. \tag{3.132}$$

Figure 3.42 shows the radiant flux density for three different temperatures for the infinitely wide bandwidth ($\lambda_1 = 0$ and $\lambda_2 = \infty$). It is seen that the radiant energy is distributed over the spectral range highly nonuniformly, with clearly pronounced maximum defined by Wien's law. A hot object radiates a significant portion of its

[18] In 1911, Wilhelm Wien (Germany, Würtzburg University) was awarded the Nobel Prize "for his discoveries regarding the laws governing the radiation of heat."

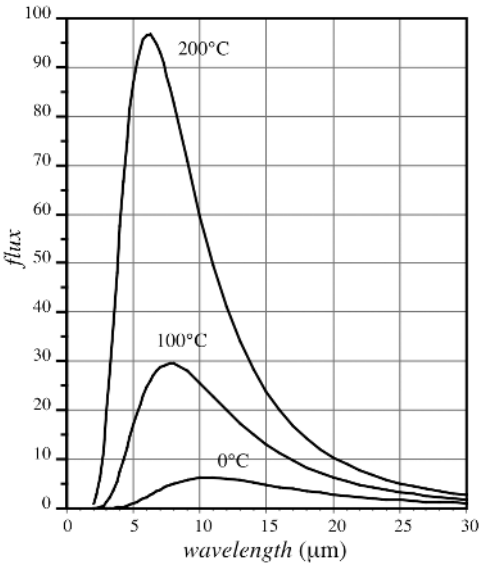


Fig. 3.42. Spectral flux density for three temperatures for the ideal radiator emanating toward infinitely cold space.

energy in the visible range, and the power radiated by the cooler objects is concentrated in the infrared and far-infrared portion of the spectrum.

Equation (3.132) is quite complex and cannot be solved analytically for any particular bandwidth. A solution can be found either numerically or by an approximation. An approximation for a broad bandwidth (when λ_1 and λ_2 embrace well over 50% of the total radiated power) is a fourth-order parabola known as the *Stefan–Boltzmann law*:

$$\Phi_{b0} = A\varepsilon\sigma T^4. \tag{3.133}$$

Here $\sigma = 5.67 \times 10^{-8}$ W/m^2 K^4 (Stefan–Boltzmann constant), A is the geometry factor, and ε is assumed to be wavelength independent [37].

Whereas wavelengths of the radiated light are temperature dependent, the magnitude of radiation is also a function of the surface property called *emissivity*, ε. Emissivity is measured on a scale from 0 to 1. It is a ratio of flux which is emanated from a surface to that emanated from the ideal emitter having the same temperature. There is a fundamental equation which relates emissivity ε, transparency γ, and reflectivity ρ:

$$\varepsilon + \gamma + \rho = 1. \tag{3.134}$$

In 1860, Kirchhoff had found that emissivity and absorptivity, α, is the same thing. As a result, for an opaque object ($\gamma = 0$), reflectivity ρ and emissivity ε are connected by a simple relationship: $\rho = 1 - \varepsilon$.

The Stefan–Boltzmann law specifies radiant power (flux) which would be emanated from a surface of temperature T toward an infinitely cold space (at absolute zero). When thermal radiation is detected by a thermal sensor,[19] the opposite radiation

[19] Here, we discuss the so-called *thermal* sensors as opposed to quantum sensors, which are described in Chapter 13.

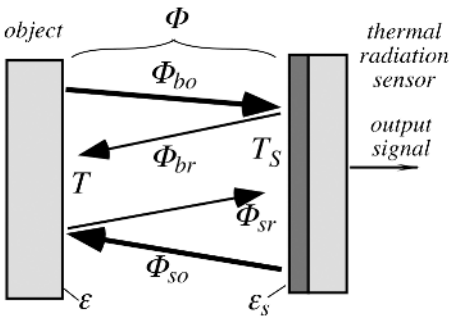


Fig. 3.43. Thermal radiation exchange between an object and a thermal radiation sensor.

from the sensor toward the object must also be taken into account. A thermal sensor is capable of responding only to a net thermal flux (i.e., flux from the object minus flux from itself). The surface of the sensor which faces the object has emissivity ε_s (and, subsequently, reflectivity $\rho_s = 1 - \varepsilon_s$). Because the sensor is only partly absorptive, not the entire flux, Φ_{b0}, is absorbed and utilized. A part of it, Φ_{ba}, is absorbed by the sensor and another part, Φ_{br}, is reflected (Fig. 3.43) back toward to object.[20] The reflected flux is proportional to the sensor's coefficient of reflectivity:

$$\Phi_{br} = -\rho_s \Phi_{b0} = -A\varepsilon(1 - \varepsilon_s)\sigma T^4. \tag{3.135}$$

A negative sign indicates an opposite direction with respect to flux Φ_{b0}. As a result, the net flux originated from the object is

$$\Phi_b = \Phi_{b0} + \Phi_{br} = A\varepsilon\varepsilon_s\sigma T^4. \tag{3.136}$$

Depending on its temperature T_s, the sensor's surface radiates its own net thermal flux toward the object in a similar way:

$$\Phi_s = -A\varepsilon\varepsilon_s\sigma T_s^4. \tag{3.137}$$

Two fluxes propagate in the opposite directions and are combined into a final net flux existing between two surfaces:

$$\Phi = \Phi_b + \Phi_s = A\varepsilon\varepsilon_s\sigma(T^4 - T_s^4). \tag{3.138}$$

This is a mathematical model of a net thermal flux which is converted by a thermal sensor into the output signal. It establishes a connection between thermal power Φ absorbed by the sensor and the absolute temperatures of the object and the sensor.

3.12.3.1 Emissivity

The emissivity of a medium is a function of its dielectric constant and, subsequently, refractive index n. The highest possible emissivity is 1. It is attributed to the so-called blackbody—an ideal emitter of electromagnetic radiation. The name implies

[20] This analysis assumes that there are no other objects in the sensor's field of view.

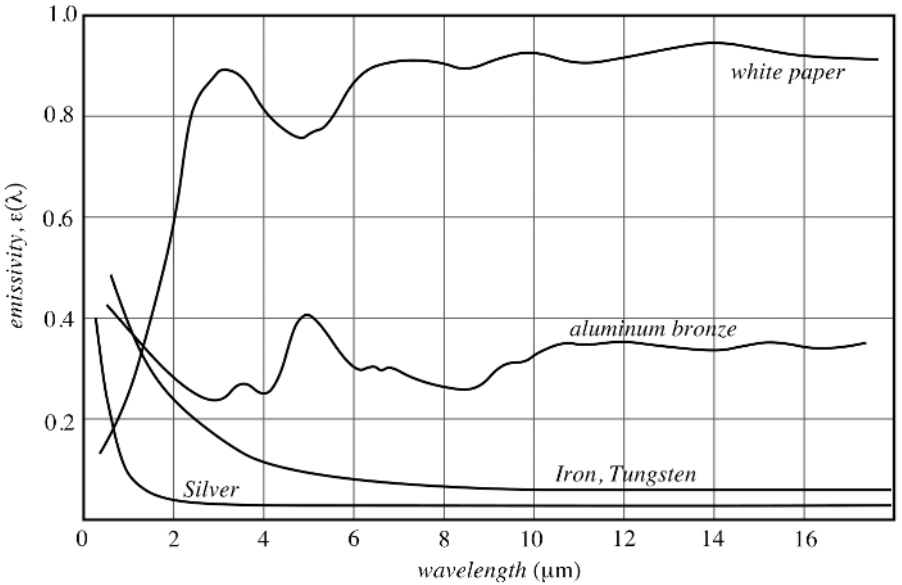


Fig. 3.44. Wavelength dependence of emissivities.

its appearance at normal room temperatures. If the object is opaque ($\gamma = 0$) and nonreflective ($\rho = 0$) according to Eq. (3.134), it becomes an ideal emitter and absorber of electromagnetic radiation (because $\varepsilon = \alpha$). It should be noted, however, that emissivity is generally wavelength dependent (Fig. 3.44). For example, a white sheet of paper is very much reflective in the visible spectral range and emits virtually no visible light. However, in the far-infrared spectral range, its reflectivity is very low and emissivity is high (about 0.92), thus making paper a good emitter of far-infrared radiation. Polyethylene, which is widely used for the fabrication of far-infrared lenses, heavily absorbs (emits) in narrow bands around 3.5, 6.8, and 13.5 μm, and is quite transparent (nonemissive) in other bands.

For many practical purposes, emissivity in a relatively narrow spectral range of thermal radiation (e.g., from 8 to 16 μm) may be considered constant. However, for precision measurements, when thermal radiation must be detected to an accuracy better than 1%, surface emissivity either must be known, or the so-called dual-band IR detectors should be employed.[21]

For a nonpolarized far-infrared light in the normal direction, emissivity may be expressed by

$$\varepsilon = \frac{4n}{(n+1)^2}. \tag{3.139}$$

[21] A dual-band detector uses two narrow spectral ranges to detect IR flux. Then, by using a ratiometric technique of signal processing, the temperature of an object is calculated. During the calculation, emissivity and other multiplicative constants are canceled out.

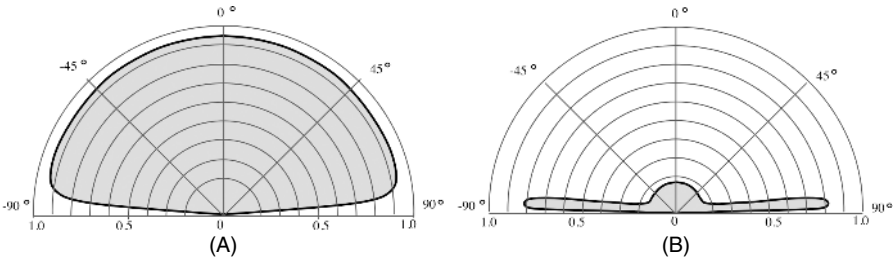


Fig. 3.45. Spatial emissivities for a nonmetal (A) and a polished metal (B).

All nonmetals are very good diffusive emitters of thermal radiation with a remarkably constant emissivity defined by Eq. (3.139) within a solid angle of about $\pm 70°$. Beyond that angle, emissivity begins to decrease rapidly to zero with the angle approaching 90°. Near 90°, emissivity is very low. A typical calculated graph of the directional emissivity of nonmetals into air is shown in Fig. 3.45A. It should be emphasized that the above considerations are applicable only to wavelengths in the far-infrared spectral range and are not true for the visible light, because emissivity of thermal radiation is a result of electromagnetic effects which occur at an appreciable depth.

Metals behave quite differently. Their emissivities greatly depend on the surface finish. Generally, polished metals are poor emitters within the solid angle of $\pm 70°$, and their emissivity increases at larger angles (Fig. 3.45B). This implies that even a very good metal mirror reflects poorly at angles approaching 90° to normal. Table A.18 in the Appendix gives typical emissivities of some materials in a temperature range between 0°C and 100°C.

Unlike most solid bodies, gases in many cases are transparent to thermal radiation. When they absorb and emit radiation, they usually do so only in certain narrow spectral bands. Some gases, such as N_2, O_2, and others of nonpolar symmetrical molecular structure, are essentially transparent at low temperatures, whereas CO_2, H_2O, and various hydrocarbon gases radiate and absorb to an appreciable extent. When infrared light enters a layer of gas, its absorption has an exponential decay profile, governed by *Beer's law*:

$$\frac{\Phi_x}{\Phi_0} = e^{-\alpha_\lambda x}, \tag{3.140}$$

where Φ_0 is the incident thermal flux, Φ_x is the flux at thickness x, and α_λ is the spectral coefficient of absorption. The above ratio is called a monochromatic transmissivity γ_λ at a specific wavelength λ. If gas is nonreflecting, then its emissivity is defined as

$$\varepsilon_\lambda = 1 - \gamma_\lambda = 1 - e^{-\alpha_\lambda x}. \tag{3.141}$$

It should be emphasized that because gases absorb only in narrow bands, emissivity and transmissivity must be specified separately for any particular wavelength. For instance, water vapor is highly absorptive at wavelengths of 1.4, 1.8, and 2.7 μm and is very transparent at 1.6, 2.2, and 4 μm.

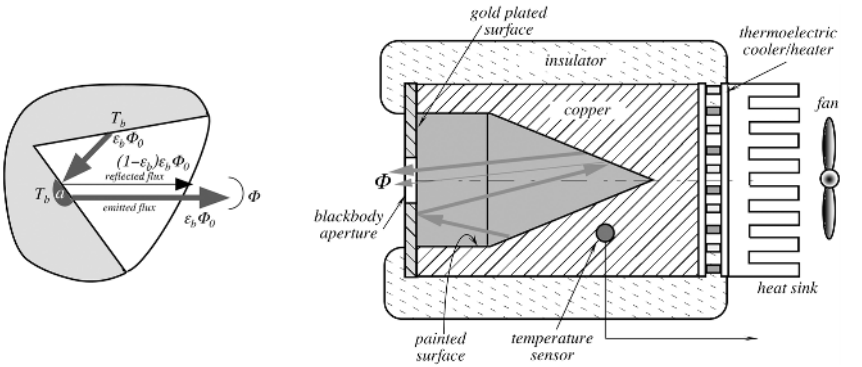


Fig. 3.46. (A) Cavity effect enhances emissivity; (B) a construction of a practical blackbody with a dual-cavity surface.

Knowing emissivity is essential when an infrared sensor is used for a noncontact temperature measurement [see Eq. (3.138)]. To calibrate such a noncontact thermometer or verify its accuracy, a laboratory standard source of heat must be constructed. The source must have precisely known emissivity and that emissivity preferably should approach unity as close as practical. A nonunity emissivity would result in reflection [Eq. (3.134)] that may introduce a significant error in detected infrared flux. There is no known material that has emissivity of 1. Thus, a practical way to artificially simulate such a surface is use of the *cavity effect*.

3.12.3.2 Cavity Effect

An interesting effect develops when electromagnetic radiation is measured from a cavity. For this purpose, a cavity means a void of a generally irregular shape inside a body whose inner wall temperature is uniform over an entire surface (Fig. 3.46A). Emissivity of a cavity opening or aperture (not of a cavity inner surface!) dramatically increases approaching unity at any wavelength, as compared with a flat surface. The cavity effect is especially pronounced when the inner walls of a void have relatively high emissivity. Let us consider a nonmetal cavity surface. All nonmetals are diffuse emitters. Also, they are diffuse reflectors. We assume that temperature and surface emissivity of the cavity are homogeneous over an entire area. An ideal object whose emissivity is equal to unity is called a *blackbody*. A blackbody would emanate from area a the infrared photon flux $\Phi_0 = a\sigma T_b^4$. However, the real object has the actual emissivity ε_b and, as a result, the flux radiated from that area is smaller: $\Phi_r = \varepsilon_b \Phi_0$. Flux which is emitted by other parts of the object toward area a is also equal to Φ_r (as the object is thermally homogeneous, we may disregard spatial distribution of flux). A substantial portion of that incident flux Φ_r is absorbed by the surface of area a,

whereas a smaller part is diffusely reflected:

$$\Phi_\rho = \rho\Phi_r = (1-\varepsilon_b)\varepsilon_b\Phi_0; \tag{3.142}$$

the combined radiated and reflected flux from area a is

$$\Phi = \Phi_r + \Phi_\rho = \varepsilon_b\Phi_0 + (1-\varepsilon_b)\varepsilon_b\Phi_0 = (2-\varepsilon_b)\varepsilon_b\Phi_0. \tag{3.143}$$

As a result, the effective emissivity may be expressed as

$$\varepsilon_e = \frac{\Phi}{\Phi_0} = (2-\varepsilon_b)\varepsilon_b \tag{3.144}$$

It follows from the above that due to a single reflection, a perceived (effective) emissivity of a cavity is equal to the surface emissivity magnified by a factor of (2–ε_b). Of course, there may be more than one reflection of radiation before it exits the cavity. In other words, the incident on area a flux could already be a result of a combined effect from the reflectance and emittance at other parts of the cavity's surface. The flux intensity will be higher than the originally emanated flux Φ_r.

For a cavity effect to work, the effective emissivity must be attributed to the cavity opening from which radiation escapes. If a sensor is inserted into the cavity too deeply facing its wall directly, blocking the reflected rays, the cavity effect may disappear and the emissivity will be equal to that of a wall surface, which is always lower.

A cavity effect will change a perceived emissivity, and if not accounted for, it may cause error in evaluation of the radiated power. To illustrate this, Fig. 3.47 shows two photographs: one is taken in visible light and the other in the mid-infrared (thermal radiation). Note that areas at the nostrils appear a little bit brighter (warmer). Yet, the temperature of the skin in these spots is the same as nearby. Two wrinkles above the mustache cause a cavity effect, which increases the skin emissivity from an average of 0.96 to a higher value. This enhances the intensity of the emanated thermal flux and gives an illusion of warmer skin.

Fabrication of a laboratory cavity blackbody is not a trivial task. For a cavity effect to work, a blackbody must have a cavity whose surface area is much larger than the exit aperture, the shape of the cavity must allow for multiple inner reflections before the flux can escape from the aperture, and the cavity wall temperature must be uniform all over its entire surface. Figure 3.46B shows an efficient way to fabricate a blackbody

Fig. 3.47. Photographs in visible light and infrared thermal radiation which is naturally emanated from the object. Note the brighter (appearing warmer) areas at the wrinkles and skin folds near the nose—a result of the cavity effect. Eyeglasses appear black (cold) because glass is opaque in the mid- and far-infrared spectral ranges and does not pass thermal radiation from the face. (Photo courtesy of Infrared Training Center, www.infraredtraining.com.)

[38] whose emissivity is over 0.999. A cavity body is fabricated of solid copper with a cavity of any shape; an inversed cone is preferable. An imbedded temperature sensor and a thermoelectric heater/cooler with a control circuit (not shown) form a thermostat that maintains the temperature of the cavity on a preset level. That may be above or below the ambient temperature. The inner portion of the cavity should be painted with organic paint. The visible color of the paint is not important, because in the infrared spectral range, there is no correlation with reflectivity in the visible range (which determines visible color). The most troublesome portion of a cavity is located near the aperture, because it is very difficult to ensure that the temperature of the left side of the blackbody (as in Fig. 3.46B) is independent of ambient and equal to the rest of the cavity walls. To minimize the effects of ambient temperature and increase the virtual cavity size, the inner surface of the front wall around the cavity is highly polished and gold plated. Thus, the front side of the cavity has very low emissivity and, thus, its temperature is not that critical. In addition, the gold surface reflects rays emitted by the right-side parts of the cavity walls that have high emissivity and thus enhances the cavity effect. The entire copper body is covered with a thermally insulating layer. It should be noted that the blackbody surface is the virtual surface of the aperture, which, in reality, is a void.

3.13 Light

Light is a very efficient form of energy for sensing a great variety of stimuli. Among many others, these include distance, motion, temperature, and chemical composition. Light has an electromagnetic nature. It may be considered a propagation of either quanta of energy or electromagnetic waves. Different portions of the wave-frequency spectrum are given special names: ultraviolet (UV), visible, near-, mid-, and far-infrared (IR), microwaves, radiowaves, and so forth. The name "light" was arbitrarily given to electromagnetic radiation which occupies wavelengths from approximately 0.1 to 100 μm. Light below the shortest wavelength that we can see (violet) is called ultraviolet, and higher than the longest that we can see (red) is called infrared. The infrared range is arbitrarily subdivided into three regions: near-infrared (from about 0.9 to 1.5 mμ), mid-infrared (1.5 to 4 μm), and far-infrared (4 to 100 μm).

Different portions of the radiation spectrum are studied by separate branches of physics. An entire electromagnetic spectrum is represented in Fig. 3.41. It spreads from γ-rays (the shortest) to radiowaves (the longest). In this section, we will briefly review those properties of light which are mostly concerned with the visible and near-infrared portions of the electromagnetic spectrum. Thermal radiation (mid- and far-infrared regions) are covered in Section 3.12.

The velocity of light c_0 in vacuum is independent of wavelengths and can be expressed as $\mu_0 = 4\pi \times 10^{-7}$ henrys/m and $\varepsilon_0 = 8.854 \times 10^{-12}$ farads/m , which are the magnetic and electric permitivities of free space:

$$c_0 = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} = 299{,}792{,}458.7 \pm 1.1 \frac{\text{m}}{\text{s}}. \tag{3.145}$$

The frequency of light waves in vacuum or any particular medium is related to its wavelength λ by Eq. (3.128), which we rewrite here as

$$\nu = \frac{c}{\lambda}, \tag{3.146}$$

where c is the speed of light in a medium.

The energy of a photon relates to its frequency as

$$E = h\nu, \tag{3.147}$$

where $h = 6.63 \times 10^{-34}$ J s (4.13×10^{-15} eV s) is Planck's constant. The energy E is measured in 1.602×10^{-19} J=1 eV (electron volt).

Ultraviolet and visible photons carry relatively large energy and are not difficult to detect. However, when the wavelength increases and moves to an infrared portion of the spectrum, the detection becomes more and more difficult. A near-infrared photon having a wavelength of 1 μm has an energy of 1.24 eV. Hence, an optical quantum detector operating in the range of 1 μm must be capable of responding to that level of energy. If we keep moving even further toward the mid- and far-infrared spectral ranges, we deal with even smaller energies. Human skin (at 37°C) radiates near- and far-infrared photons with energies near 0.13 eV, which is an order of magnitude lower than red light, making them much more difficult to detect. This is the reason why low-energy radiation is often detected by thermal detectors rather than quantum detectors.

The electromagnetic wave (now we ignore the quantum properties of light) has the additional characteristic that is *polarization* (more specifically, *plane polarization*). This means that the alternating electric field vectors are parallel to each other for all points in the wave. The magnetic field vectors are also parallel to each other, but in dealing with the polarization issues related to sensor technologies, we focus our attention on the electric field, to which most detectors of the electromagnetic radiation are sensitive. Figure 3.48A shows the polarization feature. The wave in the picture is traveling in the x-direction. It is said that the wave is polarized in the y-direction because the electric field vectors are all parallel to this axis. The plane defined by the direction of propagation (the x axis) and the direction of polarization (the y axis) is called the *plane of vibration*. In a polarized light, there are no other directions for the field vectors.

Figure 3.48B shows a randomly polarized light which is the type of light that is produced by the Sun and various incandescent light sources; however, the emerging beam in most laser configurations is polarized. If unpolarized light passes through a polarization filter (Polaroid), only specific planes can pass through and the output electric field will be as shown in Fig. 3.48C. The polarization filter transmits only those wave-train components whose electric vectors vibrate parallel to the filter direction and absorbs those that vibrate at right angles to this direction. The emerging light will be polarized according to the filter orientation. This polarizing direction in the filter is established during the manufacturing process by embedding certain long-chain molecules in a flexible plastic sheet and then stretching the sheet so that the molecules are aligned in parallel to each other. The polarizing filters are most widely used in the

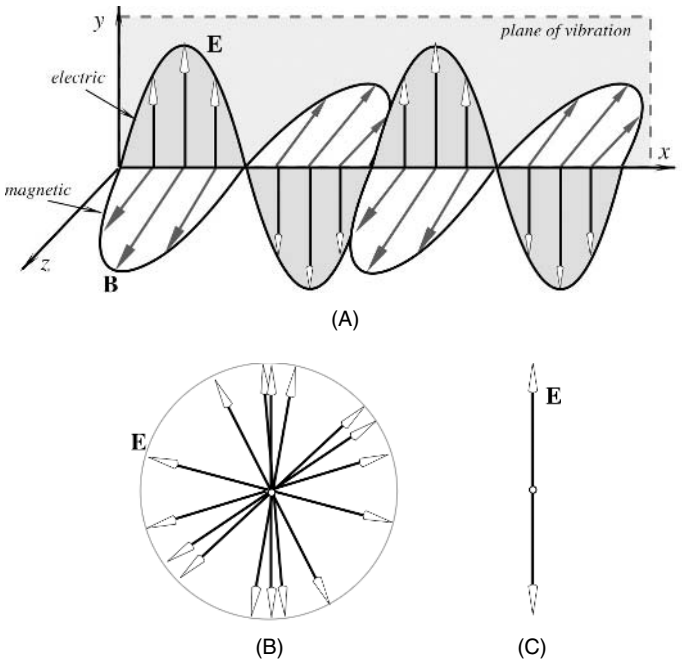


Fig. 3.48. (A) Traveling electromagnetic wave has electric and magnetic field vectors; (B) unpolarized electric field viewed along the x axis (magnetic vectors are not shown but they are always there); (C) vertically polarized electric field.

liquid-crystal displays (LCDs) and in many optical sensors that are described in the corresponding chapters of this book.

3.14 Dynamic Models of Sensor Elements

To determine a sensor's dynamic response, a variable stimulus should be applied to its input while observing the output values. Generally, a test stimulus may have any shape or form, which should be selected depending on a practical need. For instance, for determining a natural frequency of an accelerometer, sinusoidal vibrations of different frequencies are the best. On the other hand, for a thermistor probe, a step function of temperature would be preferable. In many other cases, a step or square-pulse input stimulus is often employed. The reason for that is the theoretically infinite frequency spectrum of a step function; that is, the sensor can be tested simultaneously at all frequencies.

Mathematically, a sensor can be described by a differential equation whose order depends on the sensor's physical nature and design. There are three general types of relationship between the input s and the output S: a zero-order, a first-order and a second-order response.

A *zero-order* response is a static or time independent characteristic

$$S(t) = Gs(t), \tag{3.148}$$

where G is a constant transfer function. This relationship may take any form—for instance, described by Eqs. (2.1—2.4). The important point is that G is not a function of time; that is, a zero-order response to a step function is a step function.

A *first-order* response is characterized by a first-order differential equation

$$a_1 \frac{dS(t)}{dt} + a_0 S(t) = s(t), \tag{3.149}$$

where a_1 and a_0 are constants. This equation characterizes a sensor that can store energy before dissipating it. An example of such a sensor is a temperature sensor which has a thermal capacity and is coupled to the environment through a thermal resistance. A first-order response to a step function is exponential:

$$S(t) = S_0(1 - e^{-t/\tau}), \tag{3.150}$$

where S_0 is a sensor's static response and τ is a time constant which is a measure of inertia. A typical first-order response is shown in Fig. 2.9B of Chapter 2.

A *second-order* response is characterized by a second-order differential equation

$$a_2 \frac{d^2 S(t)}{dt^2} + a_1 \frac{dS(t)}{dt} + a_0 S(t) = s(t). \tag{3.151}$$

This response is specific for a sensor or a system that contains two components which may store energy—for instance, an inductor and a capacitor, or a temperature sensor and a capacitor. A second-order response contains oscillating components and may lead to instability of the system. A typical shape of the response is shown in Fig. 2.11E of Chapter 2. A dynamic error of the second-order response depends on several factors, including its natural frequency ω_0 and damping coefficient b. A relationship between these values and the independent coefficients of Eq. (3.151) are the following:

$$\omega_0 = \sqrt{\frac{a_0}{a_2}}, \tag{3.152}$$

$$b = \frac{a_1}{2\sqrt{a_0 a_2}}. \tag{3.153}$$

A critically damped response (see Fig. 2.10 of Chapter 2) is characterized by $b = 1$. The overdamped response has $b > 1$ and the underdamped has $b < 1$. For a more detailed description of dynamic responses the reader should refer to specialized texts, (e.g., Ref. [39]).

Mathematical modeling of a sensor is a powerful tool in assessing its performance. The modeling may address two issues: static and dynamic. Static models usually deal with the sensor's transfer function as it is defined in Chapter 2. Here, we briefly outline how sensors can be evaluated dynamically. The dynamic models may have

several independent variables; however, one of them must be time. The resulting model is referred to as a lumped parameter model. In this section, mathematical models are formed by applying physical laws to some simple lumped parameter sensor elements. In other words, for the analysis, a sensor is separated into simple elements and each element is considered separately. However, once the equations describing the elements have been formulated, individual elements can be recombined to yield the mathematical model of the original sensor. The treatment is intended not to be exhaustive, but rather to introduce the topic.

3.14.1 Mechanical Elements

Dynamic mechanical elements are made of masses, or inertias, which have attached springs and dampers. Often the damping is viscous, and for the rectilinear motion, the retaining force is proportional to velocity. Similarly, for the rotational motion, the retaining force is proportional to angular velocity. Also, the force, or torque, exerted by a spring or shaft is usually proportional to displacement. The various elements and their governing equations are summarized in Table 3.4.

One of the simplest methods of producing the equations of motion is to isolate each mass or inertia and to consider it as a free body. It is then assumed that each of the free bodies is displaced from the equilibrium position, and the forces or torques acting on the body then drive it back to its equilibrium position. Newton's second law of motion can then be applied to each body to yield the required equation of motion.

For a rectilinear system, Newton's second law indicates that for a consistent system of units, *the sum of forces equals the mass times the acceleration*. In the SI system of units, force is measured in newtons (N), mass in kilograms (kg), and acceleration in meters per second squared (m/s^2).

For a rotational system, Newton's law becomes *the sum of the moments equals the moment of inertia times the angular acceleration*. The moment, or torque, has units

Table 3.4. Mechanical, Thermal, and Electrical Analogies

MECHANICAL	THERMAL	ELECTRICAL	
MASS M $F=M\frac{d(v)}{dt}$	CAPACITANCE C $Q=C\frac{dT}{dt}$	INDUCTOR L $V=L\frac{di}{dt}$	CAPACITOR C $i=C\frac{dV}{dt}$
SPRING k $F=k\int v dt$	CAPACITANCE C $T=\frac{1}{C}\int Q dt$	CAPACITOR C $V=\frac{1}{C}\int i dt$	INDUCTOR L $i=\frac{1}{L}\int V dt$
DAMPER b $F=bv$	RESISTANCE R $Q=\frac{1}{R}(T_2-T_1)$	RESISTOR R $V=Ri$	RESISTOR R $i=\frac{1}{R}V$

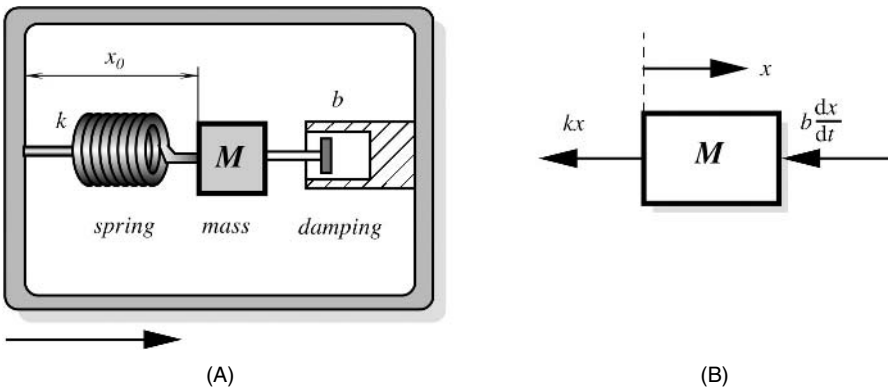


Fig. 3.49. Mechanical model of an accelerometer (A) and a free-body diagram of mass (B).

of newton meters (N m), the inertia has units of kilogram per meter squared (kg/m^2), and the angular acceleration has units of radians per second squared (rad/s^2).

Let us consider a monoaxial accelerometer, which consists of an inertia element whose movement may be transformed into an electric signal. The mechanism of conversion may be, for instance, piezoelectric. Figure 3.49A shows a general mechanical structure of such an accelerometer. The mass M is supported by a spring having stiffness k and the mass movement is damped by a damping element with a coefficient b. Mass may be displaced with respect to the accelerometer housing only in the horizontal direction. During operation, the accelerometer case is subjected to acceleration d^2y/dt^2, and the output signal is proportional to the deflection x_0 of the mass M.

Because the accelerometer mass M is constrained to linear motion, the system has one degree of freedom. Giving the mass M a displacement x from its equilibrium position produces the free-body diagram shown in Fig. 3.49B. Note that x_0 is equal to x plus some fixed displacement. Applying Newton's second law of motion gives

$$Mf = -kx - b\frac{dx}{dt}, \tag{3.154}$$

where f is the acceleration of the mass relative to the Earth and is given by

$$f = \frac{d^2x}{dt^2} - \frac{d^2y}{dt^2}. \tag{3.155}$$

Substituting for f gives the required equation of motion as

$$M\frac{d^2x}{dt^2} + b\frac{dx}{dt} + kx = M\frac{d^2y}{dt^2}. \tag{3.156}$$

Note that each term in Eq. (3.156) has units of newtons (N). The differential equation (3.156) is of a second order, which means that the accelerometer output signal may have an oscillating shape. By selecting an appropriate damping coefficient b, the output signal may be brought to a critically damped state, which, in most cases, is a desirable response.

3.14.2 Thermal Elements

Thermal elements include such things as heat sinks, heating elements, insulators, heat reflectors, and absorbers. If heat is of concern, a sensor should be regarded as a component of a larger device. In other words, heat conduction through the housing and the mounting elements, air convection, and radiative heat exchange with other objects should not be discounted.

Heat may be transferred by three mechanisms: conduction, natural and forced convection, and thermal radiation (Section 3.12). For simple lumped parameter models, the first law of thermodynamics may be used to determine the temperature changes in a body. The rate of change of a body's internal energy is equal to the flow of heat into the body minus the flow of heat out of the body, very much like fluid moves through pipes into and out of a tank. This balance may be expressed as

$$C\frac{dT}{dt} = \Delta Q, \tag{3.157}$$

where $C = Mc$ is the thermal capacity of a body (J/K), T is the temperature (K), ΔQ is the heat flow rate (W), M is the mass of the body (kg), and c is the specific heat of the material (J/kg K). The heat flow rate through a body is a function of the thermal resistance of the body. This is normally assumed to be linear, and, therefore,

$$\Delta Q = \frac{T_1 - T_2}{R}, \tag{3.158}$$

where R is the thermal resistance (K/W) and $T_1 - T_2$ is a temperature gradient across the element, where heat conduction is considered.

For illustration, we analyze a heating element (Fig. 3.50A) having temperature T_h. The element is coated with insulation. The temperature of the surrounding air is T_a. Q_1 is the rate of heat supply to the element, and Q_0 is the rate of heat loss. From Eq. (3.157),

$$C\frac{dT_h}{dt} = Q_1 - Q_0, \tag{3.159}$$

but, from Eq. (3.158),

$$Q_0 = \frac{T_h - T_a}{R}, \tag{3.160}$$

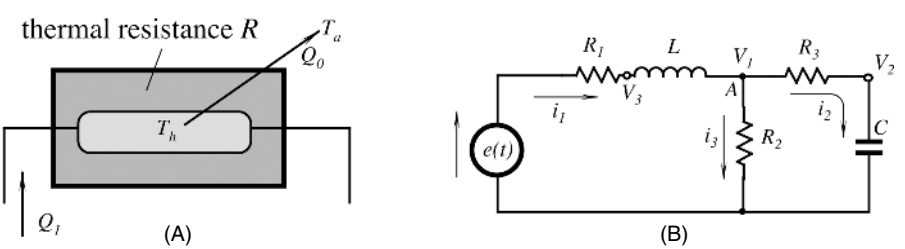


Fig. 3.50. Thermal model of a heating element (A); an electrical circuit diagram (B) with resistive, capacitive, and inductive components.

and, in the result, we obtain the differential equation

$$\frac{dT_h}{dt} + \frac{T_h}{RC} = \frac{Q_1}{C} + \frac{T_a}{RC}. \tag{3.161}$$

This is a first-order differential equation which is typical for thermal systems. A thermal element, if not a part of a control system with a feedback loop, is inherently stable. A response of a simple thermal element may be characterized by a thermal time constant, which is a product of thermal capacity and thermal resistance: $\tau_T = CR$. The time constant is measured in units of time (s) and, for a passively cooling element, is equal to the time which it takes to reach about 37% of the initial temperature gradient.

3.14.3 Electrical Elements

There are three basic electrical elements: the capacitor, the inductor, and the resistor. Again, the governing equation describing the idealized elements are given in Table 3.4. For the idealized elements, the equations describing the sensor's behavior may be obtained from Kirchhoff's laws, which directly follow from the law of conservation of energy:

Kirchhoff's first law: The total current flowing toward a junction is equal to the total current flowing from that junction (i.e., the algebraic sum of the currents flowing toward a junction is zero).

Kirchhoff's second law: In a closed circuit, the algebraic sum of the voltages across each part of the circuit is equal to the applied e.m.f.

Let us assume that we have a sensor whose elements may be represented by a circuit shown in Fig. 3.50B. To find the circuit equation, we will use the Kirchhoff's first law, which is sometimes called Kirchhoff's current law. For the node,

$$i_1 - i_2 - i_3 = 0, \tag{3.162}$$

and for each current,

$$\begin{aligned} i_1 &= \frac{e - V_3}{R_1} = \frac{1}{L}\int (V_3 - V_1)\, dt, \\ i_2 &= \frac{V_1 - V_2}{R_3} = C\frac{dV_2}{dt}, \\ i_3 &= \frac{V_1}{R_2}. \end{aligned} \tag{3.163}$$

When these expressions are substituted into Eq. (3.162), the resulting equation becomes

$$\frac{V_3}{R_1} + \frac{V_1 - V_2}{R_3} + 2\frac{V_1}{R_2} + C\frac{dV_2}{dt} - \frac{1}{L}\int (V_3 - V_1)\, dt = \frac{e}{R_1}. \tag{3.164}$$

In Equation (3.164), e/R_1 is the forcing input, and the measurable outputs are V_1, V_2, and V_3. To produce Equation (3.164), three variables i_1, i_2, and i_3 have to be

specified and three equations of motion derived. By applying the equation of constraint $i_1 - i_2 - i_3 = 0$, it has been possible to condense all three equations of motion into a single expression. Note that each element in this expression has a unit of current (A).

3.14.4 Analogies

Earlier, we considered mechanical, thermal, and electrical elements separately. However, the dynamic behavior of these systems is analogous. It is possible, for example, to take mechanical elements or thermal components, convert them into an equivalent electric circuit, and analyze the circuit using Kirchhoff's laws. Table 3.4 gives the various lumped parameters for mechanical, thermal, and electrical circuits, together with their governing equations. For the mechanical components, Newton's second law was used, and for thermal components, we apply Newton's law of cooling.

In the first column of Table 3.4 the linear mechanical elements and their equations in terms of force (F) are given. In the second column are the linear thermal elements and their equations in terms of heat (Q). In the third and fourth columns are electrical analogies (capacitor, inductor, and resistor) in terms of voltage and current (V and i). These analogies may be quite useful in a practical assessment of a sensor and for the analysis of its mechanical and thermal interface with the object and the environment.

References

1. Halliday, D. and Resnick, R. *Fundamentals of Physics*, 2nd ed. John Wiley & Sons, New York, 1986.
2. Crotzer, D.R. and Falcone, R. Method for manufacturing hygristors. U.S. patent 5,273,777; 1993.
3. Meissner, A. Über piezoelectrische Krystalle bei Hochfrequenz. Z. *Tech. Phys.* 8(74), 1927.
4. Neubert, H. K. P. Instrument transducers. An introduction to their performance and design, 2nd ed. Clarendon Press, Oxford, 1975.
5. Radice, P. F. Corona discharge poling process, U.S. patent 4,365, 283, 1982.
6. Southgate, P.D., *Appl. Phys. Lett.* 28, 250, 1976.
7. Jaffe, B., Cook, W. R., and Jaffe, H. *Piezoelectric Ceramics*. Academic Press, London, 1971.
8. Mason, W. P. *Piezoelectric Crystals and Their Application to Ultrasonics*. Van Nostrand, New York, 1950.
9. Megaw, H. D. *Ferroelectricity in Crystals*. Methuen, London, 1957.
10. Tamura, M., Yamaguchi, T., Oyaba, T., and Yoshimi, T. J. *Audio Eng. Soc.* 23(31) 1975.
11. Elliason, S. Electronic properties of piezoelectric polymers. Report TRITA-FYS 6665 from Dept. of Applied Physics, The Royal Institute of Technology, Stockholm, Sweden, 1984.
12. *Piezo Film Sensors Technical Manual*. Measurement Specialties, Inc., Fairfield, NJ, 1999; available from www.msiusa.com.

13. Oikawa, A. and Toda, K. Preparation of Pb(Zr,Ti)O3 thin films by an electron beam evaporation technique. *Appl. Phys. Lett.* 29, 491, 1976.
14. Okada, A. Some electrical and optical properties of ferroelectric lead–zirconite–lead–titanate thin films. *J. Appl. Phys.*, 48, 2905, 1977.
15. Castelano, R.N. and Feinstein, L.G. Ion-beam deposition of thin films of ferroelectric lead–zirconite–titanate (PZT). *J. Appl. Phys.*, 50, 4406, 1979.
16. Adachi, H., et al. Ferroelectric (Pb, La)(Zr, Ti)O_3 epitaxial thin films on sapphire grown by RF-planar magnetron sputtering. *J. Appl. Phys.* 60, 736, 1986.
17. Ogawa, T., Senda S., and Kasanami, T. Preparation of ferroelectric thin films by RF sputtering. *J. Appl. Phys.* 28-2, 11–14, 1989.
18. Roy, D., Krupanidhi, S. B., and Dougherty, J. Excimer laser ablated lead zirconite titanate thin films. *J. Appl. Phys.* 69, 1, 1991.
19. Yi, G., Wu, Z., and Sayer, M. Preparation of PZT thin film by sol-gel processing: electrical, optical, and electro-optic properties. *J. Appl. Phys.* 64(5), 2717–2724, 1988.
20. Kawai, H. The piezoelectricity of poly(vinylidene fluoride). *Jpn. J. of Appl. Phys.* 8, 975–976, 1969.
21. Meixner, H., Mader, G., and Kleinschmidt, P. Infrared sensors based on the pyroelectric polymer polyvinylidene fluoride (PVDF). *Siemens Forsch. Entwicl. Ber. Bd.* 15(3), 105–114, 1986.
22. Kleinschmidt, P. Piezo- und pyroelektrische Effekte. In: *Sensorik*. W. Heyward. Springer, Heidelberg, 1984, Chap. 6.
23. *Semiconductor Sensors. Data Handbook.* Philips Export B.V, Eindhoven, 1988.
24. Ye, C., Tamagawa, T., and Polla, D.L. Pyroelectric $PbTiO_3$ thin films for microsensor applications. In: *Transducers'91. International conference on Solid-State Sensors and Actuators. Digest of Technical Papers*, Schooley, J., ed. IEEE, New York, 1991, pp. 904–907.
25. Beer, A. C. *Galvanomagnetic Effect in Semiconductors. Solid State Physics.* F. Seitz and D. Turnbull, eds. Academic Press, New York, 1963.
26. Putlye, E. H. *The Hall Effect and Related Phenomena*. Semiconductor monographs., Hogarth, ed. Butterworths, London, 1960.
27. Sprague *Hall Effect and Optoelectronic Sensors. Data Book* SN-500, 1987.
28. Williams, J. Thermocouple measurement, In: *Linear applications handbook*, Linear Technology Corp., 1990.
29. Seebeck, T., Dr. Magnetische Polarisation der Metalle und Erze durch TemperaturDifferenz. *Abhaandulgen der Preussischen Akademic der Wissenschaften*, pp. 265–373, 1822–1823.
30. Benedict, R. P. *Fundamentals of Temperature, Pressure, and Flow Measurements*, 3rd ed. John Wiley & Sons, New York, 1984.
31. LeChatelier, *H. Copt. Tend.*, 102, 1886.
32. Carter, E. F. ed., *Dictionary of Inventions and Discoveries*. Crane, Russak and Co., New York, 1966.
33. Peltier, J.C.A. Investigation of the heat developed by electric currents in homogeneous materials and at the junction of two different conductors, *Ann. Phys. Chem.*, 56, 1834.

34. Thomson, W. On the thermal effects of electric currents in unequal heated conductors. *Proc. R. Soc.* VII, 1854.
35. *Manual on the Use of Thermocouples in Temperature Measurement*. ASTM, Philadelphia, 1981.
36. Doebelin, E.O. *Measurement Systems: Application and Design*, 4th ed. McGraw-Hill, New York, 1990.
37. Holman J. P. *Heat Transfer*, 3rd ed. McGraw-Hill, New York, 1972.
38. Fraden, J. Blackbody cavity for calibration of infrared thermometers. U.S. patent 6447160, 2002.
39. Thompson, S. *Control Systems. Engineering & Design*. Longman Scientific & Technical, Essex, UK, 1989.
40. MacDonald, D.K.C. Thermoelectricity: an introduction to the principles. John Wiley & Sons, New York, 1962.

This page intentionally left blank

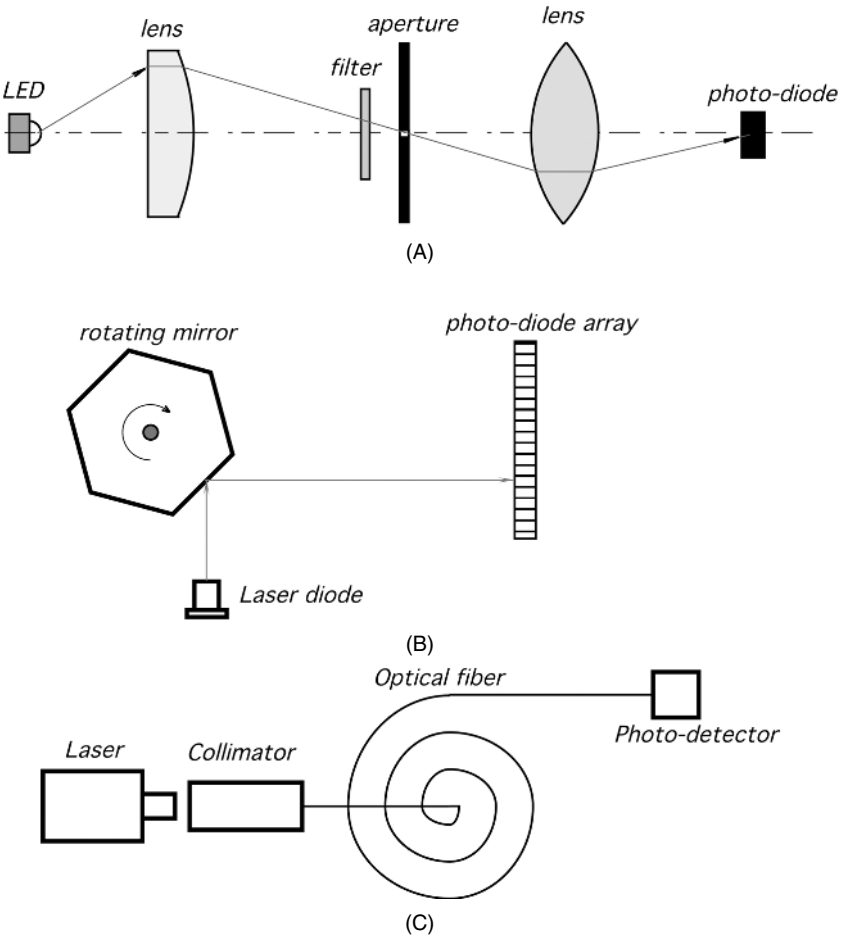


Fig. 4.1. Examples of optical systems that use refraction (A) and reflection (B,C).

filaments in the electric bulbs, light-emitting diodes (LEDs), gas-discharge lamps, lasers, laser diodes, heaters, and coolers.

Figure 4.1 shows several examples of the manipulation of light in sensors. Most of these methods involve changing the direction of light; others use a selective blocking of certain wavelengths. The latter is called filtering (filter in Fig. 4.1A). The light direction can be changed by use of the physical effect of *reflection* with the help of mirrors, diffractive gratings, optical waveguides, and fibers. Also, the light direction can be changed by *refraction* with the help of lenses, prisms, windows, chemical solutions, crystals, organic materials, and biological objects. While passing through these objects, properties of the light may be modified (modulated) by a measured stimulus. Then, the task of a sensor designer is to measure the degree of such modulation and relate it to the stimulus. What can be modulated in light? The intensity, direction of

propagation, polarization, spectral contents of light beam can all be modified, and even the speed of light and phase of its wavelength can be changed.

4.1 Radiometry

Let us consider light traveling through a three-layer material. All layers are made of different substances called media. Figure 4.2 shows what happens to a ray of light which travels from the first medium into a flat plate of a second medium, and then to a third medium. Part of the incident light is reflected from a planar boundary between the first and second media according to the *law of reflection*, which historically is attributed to Euclid:

$$\Theta_1 = \Theta'_1 \tag{4.1}$$

A part of light enters the plate (Medium 2) at a different angle. The new angle Θ_2 is governed by the *refraction law*, which was discovered in 1621 by Willebrord Snell (1580–1626) and is known as *Snell's law*:

$$n_1 \sin \Theta_1 = n_2 \sin \Theta_2, \tag{4.2}$$

where n_1 and n_2 are the indices of refraction of two media.

In any medium, light moves slower than in vacuum. An *index of refraction* is a ratio of velocity of light in vacuum, c_0, to that in a medium, c:

$$n = \frac{c_0}{c}, \tag{4.3}$$

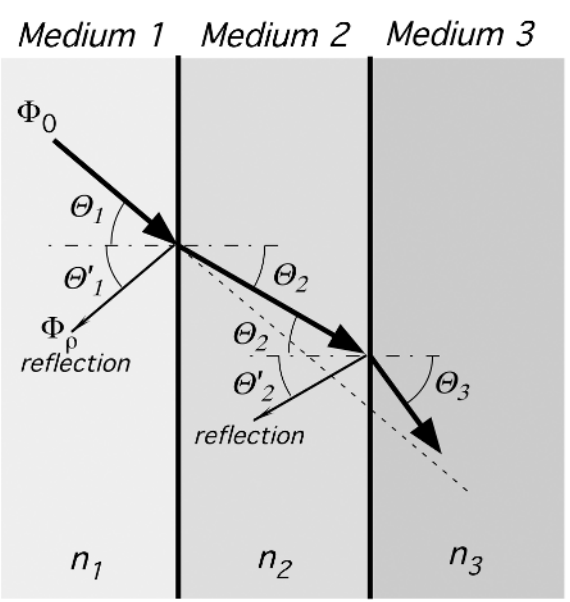


Fig. 4.2. Light passing through materials with different refractive indices.

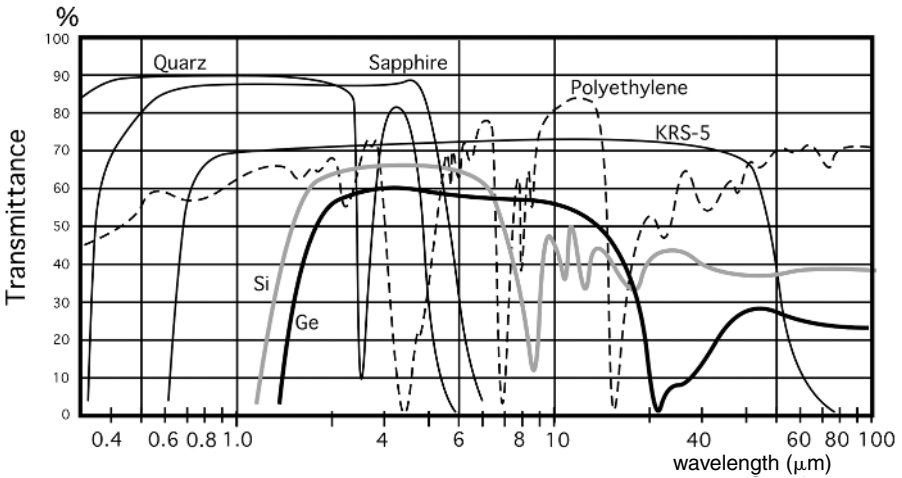


Fig. 4.3. Transparency characteristics for various optical materials.

Because $c < c_0$, the refractive index of a medium is always more than unity. The velocity of light in a medium directly relates to a dielectric constant ε_r of a medium, which subsequently determines the refractive index:

$$n = \sqrt{\varepsilon_\mathrm{r}}. \tag{4.4}$$

Generally, n is a function of wavelength. The wavelength dependence of the index of refraction is manifested in a prism, which was used by Sir Isaac Newton in his experiments with the light spectrum. In the visible range, the index of refraction n is often specified at a wavelength of 0.58756 μm, the yellow-orange helium line. Indices of refraction for some materials are presented in Table A.19 in the Appendix.

A refractive index dependence of wavelengths is called a dispersion. The change in n with the wavelength is usually very gradual, and often negligible, unless the wavelength approaches a region where the material is not transparent. Figure 4.3 shows transparency curves of some optical materials.

A portion of light reflected from a boundary at angle Θ_1', depends on light velocities in two adjacent media. The amount of reflected flux Φ_ρ relates to incident flux Φ_0 through the *coefficient of reflection* ρ, which can be expressed by means of refractive indices:

$$\rho = \frac{\Phi_\rho}{\Phi_0} = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2. \tag{4.5}$$

Equations (3.139) and (4.5) indicate that both the reflection and the absorption (emissivity) depend solely on the refractive index of the material at a particular wavelength.

If the light flux enters from air into an object having refractive index n, Eq. (4.5) is simplified to

$$\rho = \left(\frac{n-1}{n+1}\right)^2. \tag{4.6}$$

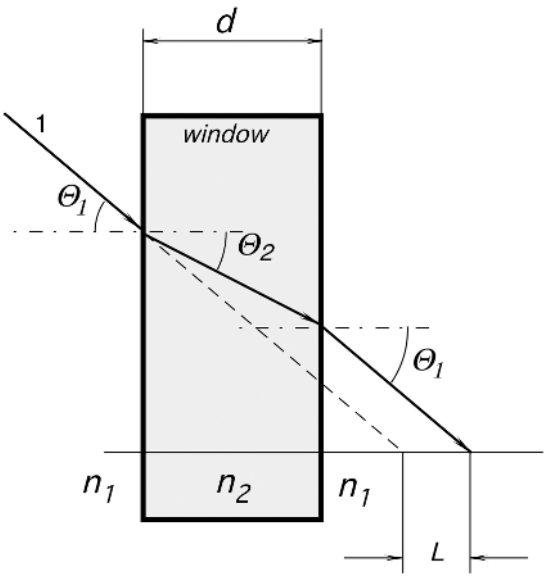


Fig. 4.4. Light passing through an optical plate.

Before light exits the second medium (Fig. 4.2) and enters the third medium having refractive index n_3, another part of it is reflected internally from the second boundary between the n_2 and n_3 media at angle Θ_2'. The remaining portion of light exits at angle Θ_3, which is also governed by Snell's law. If media 1 and 3 are the same (e.g., air) at both sides of the plate, then $n_1 = n_3$ and $\Theta_1 = \Theta_3$. This case is illustrated in Fig. 4.4. It follows from Eq. 4.5 that the coefficients of reflection are the same for light striking a boundary from either direction—approaching from the higher or lower index of refraction.

A combined coefficient of two reflections from both surfaces of a plate can be found from a simplified formula:

$$\rho_2 \approx \rho_1(2 - \rho_1), \tag{4.7}$$

where ρ_1 is the reflective coefficient from one surface. In reality, the light reflected from the second boundary is reflected again from the first boundary back to the second boundary, and so on. Thus, assuming that there is no absorption in the material, the total reflective loss within the plate can be calculated through the refractive index of the material:

$$\rho_2 = 1 - \frac{2n}{n^2 + 1}. \tag{4.8}$$

The reflection increases for higher differences in refractive indices. For instance, if visible light travels without absorption from air through a heavy flint glass plate, two reflectances result in a loss of about 11%, whereas for the air–germanium–air interfaces (in the far-infrared spectral range), the reflective loss is about 59%. To

reduce losses, optical materials are often given antireflective coatings, which have refractive indices and thickness geared to specific wavelengths.

The radiant energy balance Eq. (3.134) should be modified to account for two reflections in an optical material:

$$\rho_2 + \alpha + \gamma = 1, \tag{4.9}$$

where α is a coefficient of absorption and γ is a coefficient of transmittance. In a transparency region, $\alpha \approx 0$, therefore, transmittance is:

$$\gamma = 1 - \rho_z \approx \frac{2n}{n^2 + 1}. \tag{4.10}$$

Equation (4.10) specifies the maximum theoretically possible transmittance of the optical plate. In the above example, transmittance of a glass plate is 88.6% (visible), whereas transmittance of a germanium plate is 41% (far IR). In the visible range, germanium transmittance is zero, which means that 100% of light is reflected and absorbed. Figure 4.5 shows reflectance and transmittance of a thin plate as functions of refractive indices. Here, a plate means any optical device (like a window or a lens) operating within its useful spectral range, [i.e., where its absorptive loss is small ($\alpha \approx 0$)].

Figure 4.6 shows an energy distribution within an optical plate when incident light flux Φ_0 strikes its surface. A part of incident flux, Φ_ρ, is reflected, another part, Φ_α, is absorbed by the material, and the third part, Φ_γ, is transmitted through. The

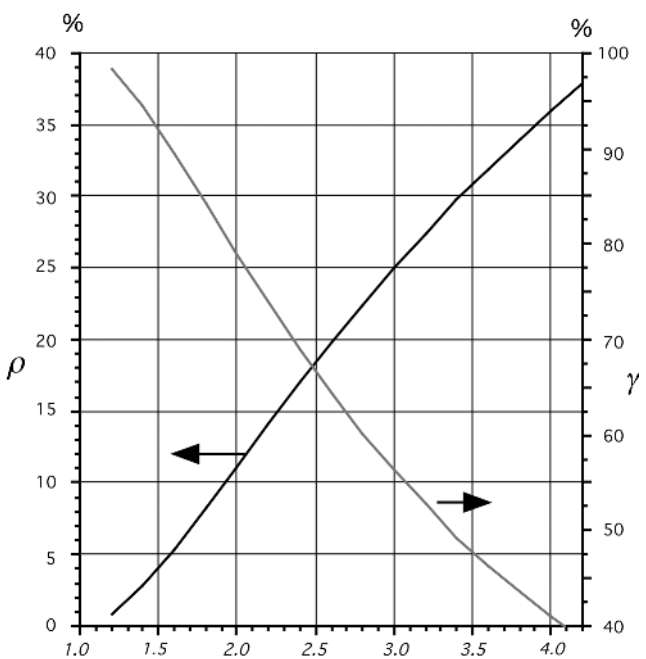


Fig. 4.5. Reflectance and transmittance of a thin plate as functions of a refractive index.

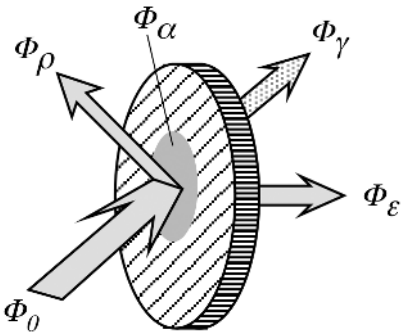


Fig. 4.6. Radiant energy distribution at an optical plate.

absorbed portion of light is converted into heat, a portion of which ΔP is lost to a supporting structure and surroundings through thermal conduction and convection. The rest of the absorbed light raises the temperature of the material. The temperature increase may be of concern when the material is used as a window in a powerful laser. Another application where temperature increase may cause problems is in far-infrared detectors. The problem is associated with the flux $\Phi_\varepsilon = \Phi_\alpha - \Delta P$, which is radiated by the material due to its temperature change. This is called a secondary radiation. Naturally, a radiated spectrum relates to a temperature of the material and is situated in the far-infrared region of the spectrum. The spectral distribution of the secondary radiation corresponds to the absorption distribution of the material because absorptivity and emissivity are the same thing.

For materials with low absorption, the absorption coefficient can be determined through a temperature rise in the material:

$$\alpha = \frac{mc}{\Phi\gamma}\frac{2n}{n^2+1}\left(\frac{dT_g}{dt} + \frac{dT_L}{dt}\right)T_0, \tag{4.11}$$

where m and c are the mass and the specific heat of the optical material, respectively, and T_g and T_L are the slopes of the rising and lowering parts of the temperature curve of the material, respectively, at test temperature T_0. Strictly speaking, light in the material is lost not only due to absorption but to scattering as well. A combined loss within material depends on its thickness and can be expressed through the so-called *attenuation coefficient* g and the thickness of the sample h. The transmission coefficient can be determined from Eq. (4.10), which is modified to account for the attenuation:

$$\gamma \approx (1-\rho_2)e^{-gh}. \tag{4.12}$$

The attenuation (or extinction) coefficient g is usually specified by manufacturers of optical materials.

4.2 Photometry

When using light-sensitive devices (photodetectors), it is critical to take into a consideration both the sensor and the light source. In some applications, light is received from independent sources; in others, the light source is a part of the measurements

system. In any event, the so-called photometric characteristics of the optical system should be taken into account. Such characteristics include light, emittance, luminance, brightness, and so forth.

To measure radiant intensity and brightness, special units have been devised. Radiant flux (energy emitted per unit time), which is situated in a visible portion of the spectrum, is referred to as luminous flux. This distinction is due to the inability of the human eye to respond equally to like power levels of different visible wavelengths. For instance, one red and one blue light of the same intensity will produce very different sensations; the red will be perceived as much brighter. Hence, comparing lights of different colors, the watt becomes a poor measure of brightness and a special unit called a *lumen* was introduced. It is based on a standard radiation source with molten platinum formed in a shape of a blackbody and visible through a specified aperture within a solid angle of one steradian. A solid angle is defined in a spherical geometry as

$$\omega = \frac{A}{r^2}, \tag{4.13}$$

where r is the spherical radius and A is the spherical surface of interest. When $A = r$, the unit is called a spherical radian or *steradian* (sr) (see Table 1.7).

Illuminance is given as

$$E = \frac{dF}{dA}; \tag{4.14}$$

that is, a differential amount of luminous flux (F) over a differential area. It is most often expressed in lumens per square meter (square foot), or foot-meter (foot-candle). The luminous intensity specifies flux over solid angle:

$$I_L = \frac{dF}{d\omega}; \tag{4.15}$$

most often, it is expressed in lumens per steradian or candela. If the luminous intensity is constant with respect to the angle of emission, Eq. (4.15) becomes

$$I_L = \frac{F}{\omega}. \tag{4.16}$$

If the wavelength of the radiation varies but the illumination is held constant, the radiative power in watts is found to vary. A relationship between illumination and radiative power must be specified at a particular frequency. The point of specification has been taken to be at a wavelength of 0.555 μm, which is the peak of the spectral response of the human eye. At this wavelength, 1 W of radiative power is equivalent to 680 lumens. For the convenience of the reader, some useful terminology is given in Table 4.1.

In the selection of electro-optical sensors, design considerations of light sources are of prime concern. A light source will effectively appear as either a *point source* or as an *area source*, depending on the relationship between the size of the source and the distance between the source and the detector. Point sources are arbitrarily defined as those whose diameter is less than 10% of the distance between the source

Table 4.1. Radiometric and Photometric Terminology

Description	Radiometric	Photometric
Total flux	Radiant flux (F) in W	Luminous flux (F) in lumens
Emitted flux density at a source surface	Radiant emittance (W) in W/cm^2	Luminous emittance (L) in lumens/cm^2 (lamberts) or lumens/ft^2 (foot-lamberts)
Source intensity (point source)	Radiant intensity (I_r) in W/sr	Luminous intensity (I_L) in lumens/sr (candela)
Source intensity (area source)	Radiance (B_r) in W/sr/cm^2	Luminance (B_L) in lumens/sr/cm^2 (lambert)
Flux density incident on a receiver surface	Irradiance (H) in W/cm^2	Illuminance (E) in lumens/cm^2 (candle) or lumens/ft^2 (foot-candle)

Source: Adapted from Ref. [2].

and the detector. Although it is usually desirable that a photodetector be aligned such that its surface area is tangent to the sphere with the point source at its center, it is possible that the plane of the detector can be inclined from the tangent plane. Under this condition, the incident flux density (irradiance) is proportional to the cosine of the inclination angle φ:

$$H = \frac{I_r}{\cos\varphi}, \tag{4.17}$$

and the illuminance,

$$E = \frac{I_L}{r^2}\cos^{\varphi}. \tag{4.18}$$

The area sources are arbitrarily defined as those whose diameter is greater than 10% of the separation distance. A special case that deserves some consideration occurs when the radius R of the light source is much larger than the distance r to the sensor. Under this condition,

$$H = \frac{B_r A_s}{r^2 + R^2} \approx \frac{B_r A_s}{R^2}, \tag{4.19}$$

where A_s is the area of the light source and B_r is the radiance. Because the area of the source $A_s = \pi R^2$, irradiance is

$$H \approx B_r \pi = W; \tag{4.20}$$

that is, the emitted and incident flux densities are equal. If the area of the detector is the same as the area of the source and $R \gg r$, the total incident energy is approximately the same as the total radiated energy, (i.e., unity coupling exists between the source and the detector). When the optical system is comprised of channeling, collimating, or focusing components, its efficiency and, subsequently, coupling coefficient must be considered. Important relationships for point and area light sources are given in Tables 4.2 and 4.3.

Table 4.2. Point Source Relationships

Description	Radiometric	Photometric
Point source intensity	I_r, W/sr	I_L, lumens/sr
Incident flux density	Irradiance, $H = I_r/r^2$, W/m^2	illuminance, $E = I_L/r^2$, lumens/m^2
Total flux output of a point source	$P = 4\pi I_r$, watts	$F = 4\pi I_L$, lumens

Source: Adapted from Ref. [2].

Table 4.3. Area Source Relationships

Description	Radiometric	Photometric
Point source intensity	B_r, W/(cm^2 sr)	B_L, lumens/(cm^2 sr)
Emitted flux density	$W = \pi B_r$, W/cm^2	$L = \pi B_L$, lumens/cm^2
Incident flux density	$H = \dfrac{B_r A_s}{r^2 + R^2}$, W/cm^2	$E = \dfrac{B_L A_s}{r^2 + R^2}$, lumens/cm^2

Source: Adapted from Ref. [2].

4.3 Windows

The main purpose of windows is to protect interiors of sensors and detectors from the environment. A good window should transmit light rays in a specific wavelength range with minimal distortions. Therefore, windows should possess appropriate characteristics depending on a particular application. For instance, if an optical detector operates under water, its window should possess the following properties: a mechanical strength to withstand water pressure, a low water absorption, a transmission band corresponding to the wavelength of interest, and an appropriate refractive index which preferably should be close to that of water. A useful window which can withstand high pressures is spherical, as shown in Fig. 4.7. To minimize optical distortions, two limitations should be applied to a spherical window: an aperture D (its largest dimension) must be smaller than the window's spherical radius R_1, and a thickness d of the window must be uniform and much smaller than radius R_1. If these conditions are not met, the window becomes a concentric spherical lens.

A surface reflectivity of a window must be considered for its overall performance. To minimize a reflective loss, windows may be given special antireflective coatings (ARCs) which may be applied on either one or both sides of the window. These are the coatings which give bluish and amber appearances to popular photographic lenses and filters. Due to a refraction in the window (see Fig. 4.4), a passing ray is shifted by a distance L which, for small angles Θ_1, may be found from the formula

$$L = d\frac{n-1}{n}, \tag{4.21}$$

where n is the refractive index of the material.

Sensors operating in the far-infrared range require special windows which are opaque in the visible and ultraviolet spectral regions and quite transparent in the wavelength of interest. Several materials are available for the fabrication of such

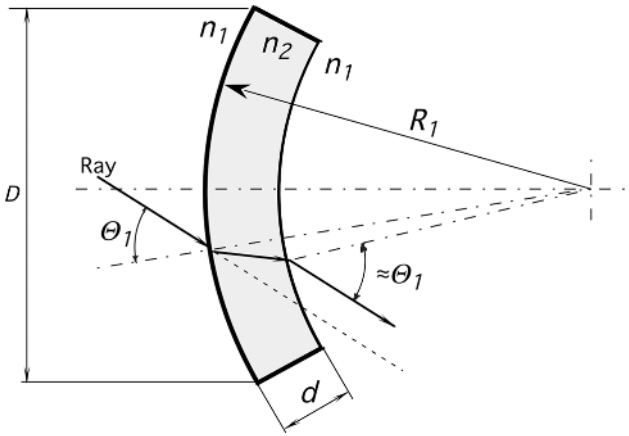


Fig. 4.7. Spherical window.

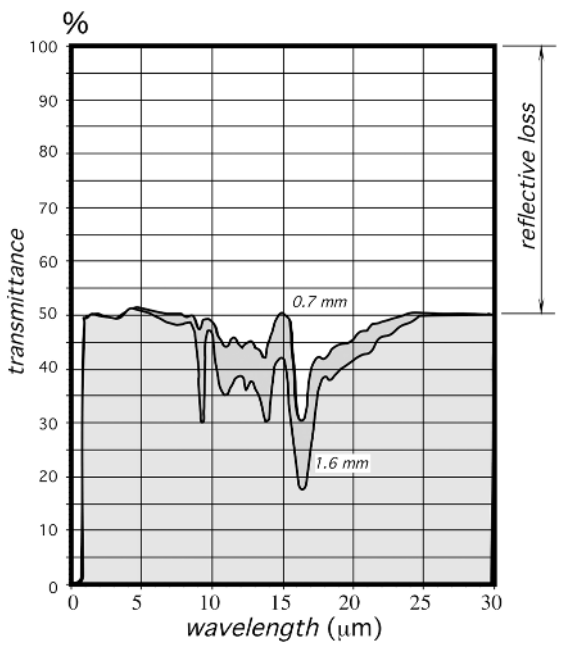


Fig. 4.8. Spectral transmittance of a silicon window. Note that the majority of loss is due to a reflection from two surfaces.

windows. Spectral transmittances of some materials are shown in Fig. 4.3. When selecting the material for a far-infrared window, the refractive index must be seriously considered because it determines the coefficient of reflectivity, absorptivity, and, eventually, transmittance. Figure 4.8 shows spectral transmittances of two silicon windows having different thicknesses. The total radiation (100%) at the window is divided into three portions: reflected (about 50% over the entire spectral range),

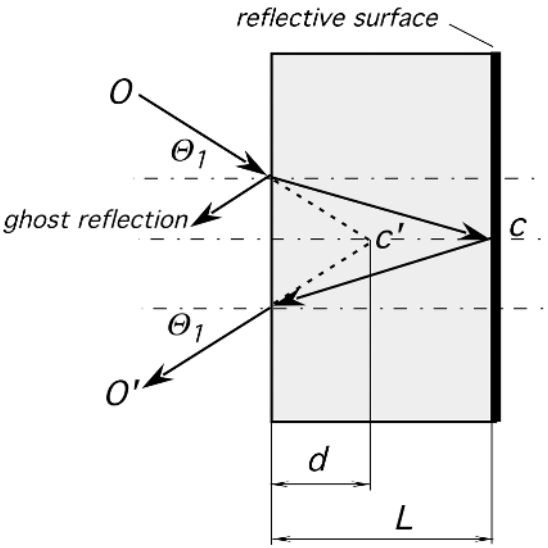


Fig. 4.9. Second surface mirror.

absorptive (varies at different wavelengths), and transmitted, which is whatever is left after reflection and absorption. Because all windows are characterized by specific spectral transmissions, often they are called *filters*.

4.4 Mirrors

A mirror is the oldest optical instrument ever used or designed. Whenever light passes from one medium to another, there is some reflection. To enhance a reflectivity, a single or multilayer reflecting coating is applied either on the front (first surface) or the rear (second surface) of a plane-parallel plate or other substrate of any desirable shape. The first surface mirrors are the most accurate. In the second surface mirror, light must enter a plate having, generally, a different index of refraction than the outside medium.

Several effects in the second surface mirror must be taken into consideration. First, due to the refractive index n of a plate, a reflective surface appears closer (Fig. 4.9). The virtual thickness d of the carrier for smaller angles Θ_1 may be found from a simple formula:

$$d \approx \frac{L}{n}. \tag{4.22}$$

The front surface of the second surface mirror may also reflect a substantial amount of light, creating the so-called ghost reflection. For instance, a glass plate reflects about 4% of visible light. Further, a carrier material may have a substantial absorption in the wavelength of interest. For instance, if a mirror operates in a far-infrared spectral range, it should use either first surface metallization or a second

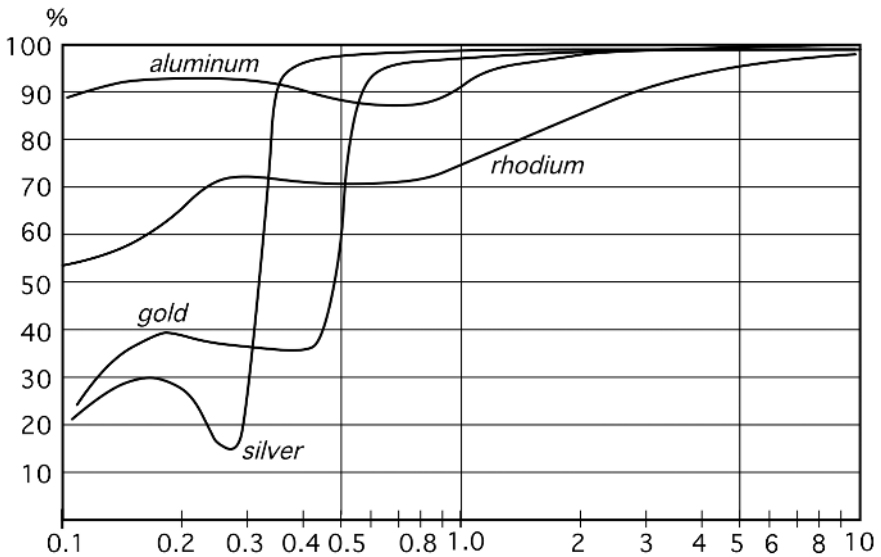


Fig. 4.10. Spectral reflectances of some mirror coatings.

surface where the substrate is fabricated of ZnSe or other long-wavelength transparent materials. Materials such as Si or Ge have too strong a surface reflectivity to be useful for the fabrication of the second surface mirrors.

Reflecting coatings applied to a surface for operation in the visible and near-infrared ranges can be silver, aluminum, chromium, and rhodium. Gold is preferable for the far-infrared spectral range devices. By selecting an appropriate coating, the reflectance may be achieved of any desired value from 0 to 1 (Fig. 4.10).

The best mirrors for broadband use have pure metallic layers, vacuum-deposited or electrolytically deposited on glass, fused silica, or metal substrates. Before the reflective layer deposition, to achieve a leveling effect a mirror may be given an undercoat of copper, zirconium–copper, or molybdenum.

Another useful reflector which may serve as a second surface mirror without the need for reflective coatings is a prism, where the effect of *total internal reflection* (TIR) is used. The angle of a total internal reflection is a function of the refractive index:

$$\Theta_0 = \arcsin\left(\frac{1}{n}\right). \tag{4.23}$$

The total internal reflectors are the most efficient in the visible and near-infrared spectral ranges, as the reflectivity coefficient is close to unity. The TIR principle is fundamental for the operation of the optical fibers.

A reflective surface may be formed practically in any shape to divert the direction of light travel. In the optical systems, curved mirrors produce effects equivalent to that of lenses. The advantages they offer include (1) higher transmission, especially in the longer-wavelength spectral range, where lenses become less efficient due to

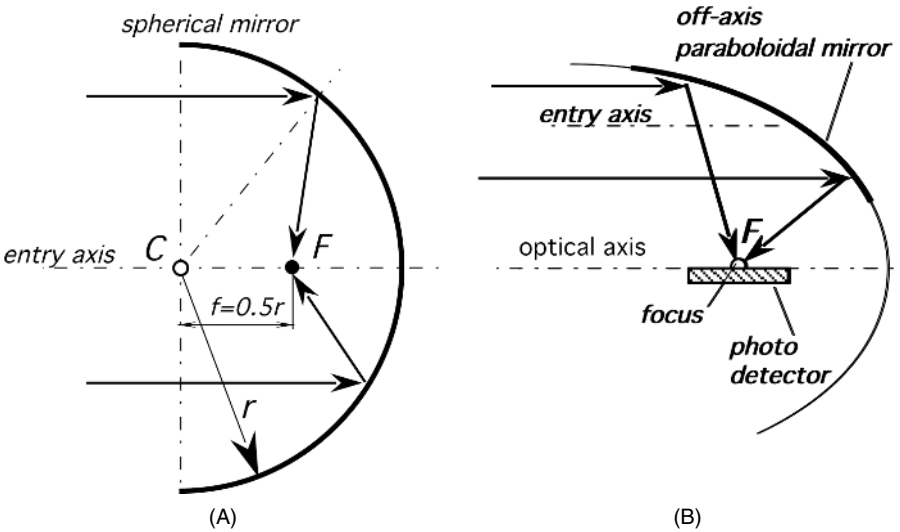


Fig. 4.11. Spherical (A) and parabolic (B) first surface mirrors.

higher absorption and reflectance loss, (2) absence of distortions incurred by refracting surfaces due to dispersion (chromatic aberrations), and (3) lower size and weight as compared with many types of lenses. Spherical mirrors are used whenever light must be collected and focused (*focus* is from the Latin meaning *fireplace*—a gathering place in a house). However, spherical mirrors are good only for the parallel or near-parallel beams of light that strike a mirror close to normal. These mirrors suffer from imaging defects called aberrations. Figure 4.11A shows a spherical mirror with the center of curvature in point C. A focal point is located at a distance of one-half of the radius from the mirror surface. A spherical mirror is astigmatic, which means that the off-axis rays are focused away from its focal point. Nevertheless, such mirrors prove very useful in detectors where no quality imaging is required—for instance, in infrared motion detectors, which are covered in detail in Section 6.5 of Chapter 6.

A parabolic mirror is quite useful for focusing light off-axis. When it is used in this way, there is complete access to the focal region without shadowing, as shown in Fig. 4.11B.

4.5 Lenses

Lenses[1] are useful in sensors and detectors to divert the direction of light rays and arrange them in a desirable fashion. Figure 4.12 shows a plano-convex lens, which has one surface spherical and the other flat. The lens has two focuses at both sides: F and F', which are positioned at equal distances $-f$ and f from the lens. When light rays from object G enters the lens, their directions change according to Snell's law.

[1] The word *lens* is from the Latin name for lentils. A lentil seed is flat and round, and its sides bulge outward—just like a convex lens.

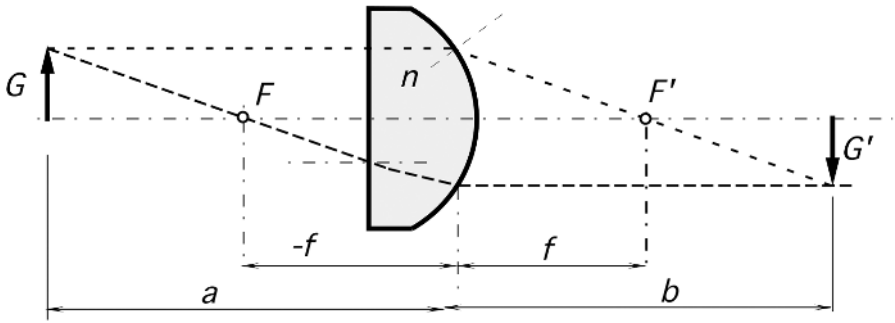


Fig. 4.12. Geometry of a plano-convex lens.

To determine the size and the position of an image created by the lens, it is convenient to draw two rays that have special properties. One is parallel to the optical axis, which is a line passing through the sphere's center of curvature. After exiting the lens, that ray goes through focus F'. The other ray first goes though focus F, and upon exiting the lens, it propagates in parallel with the optical axis. A thin lens whose radius of curvature is much larger than thickness of the lens has a focal distance f, which may be found from the equation

$$\frac{1}{f} = (n-1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right), \tag{4.24}$$

where r_1 and r_2 are radii of lens curvature. Image G' is inverted and positioned at a distance b from the lens. That distance may be found from a thin-lens equation:

$$\frac{1}{f} = \frac{1}{a} + \frac{1}{b}. \tag{4.25}$$

For the thick lenses where thickness t is comparable to the radii of curvature, a focal distance may be found from

$$f = \frac{nr_1r_2}{(n-1)|n(r_1+r_2) - t(n-1)|}. \tag{4.26}$$

Several lenses may be combined into a more complex system. For two lenses separated by a distance d, a combination focal length may be found from

$$f = \frac{f_1 f_2}{f_1 + f_2 - d}. \tag{4.27}$$

4.6 Fresnel Lenses

Fresnel lenses are optical elements with step-profiled surfaces. They prove to be very useful in sensors and detectors where a high quality of focusing is not required. Major applications include light condensers, magnifiers, and focusing element in occupancy

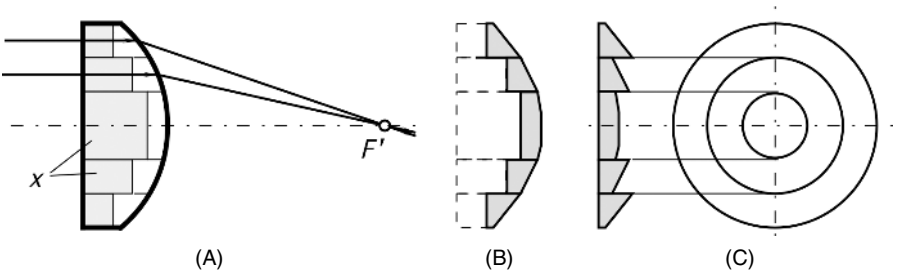


Fig. 4.13. Concept of a Fresnel lens.

detectors. Fresnel lenses may be fabricated of glass, acrylic (visible and near-infrared range), or polyethylene (far-infrared range). The history of Fresnel lenses began in 1748, when Count Buffon proposed grinding out a solid piece of glass lens in steps of the concentric zones in order to reduce the thickness of the lens to a minimum and to lower energy loss. He realized that only the surface of a lens is needed to refract light, because once the light is inside the lens, it travels in a straight line. His idea was modified in 1822 by Augustin Fresnel (1788–1827), who constructed a lens in which the centers of curvature of the different rings receded from the axis according to their distances from the center, so as to practically eliminate spherical aberration.

The concept of that lens is illustrated in Fig. 4.13, where a regular plano-convex lens is depicted. The lens is sliced into several concentric rings. After slicing, all rings still remain lenses which refract incident rays into a common focus defined by Eq. (4.24). A change in an angle occurs when a ray exits a curved surface. The section of a ring marked by the letter x does not contribute to the focusing properties. If all such sections are removed, the lens will look like it is shown in Fig. 4.13B and will fully retain its ability to focus light rays. Now, all of the rings may be shifted with respect to one another to align their flat surfaces (Fig. 4.13C). A resulting near-flat lens is called Fresnel, which has nearly the same focusing properties as the original plano-convex lens. A Fresnel lens basically consists of a series of concentric prismatic grooves, designed to cooperatively direct incident light rays into a common focus.

The Fresnel lens has several advantages over a conventional lens, such as low weight, thin size, ability to be curved (for a plastic lens) to any desirable shape, and, most importantly, lower absorption loss of light flux. The last feature is very important for the fabrication of mid- and far-infrared lenses where absorption in the material may be significant. This is the reason why polymer Fresnel lenses are used almost exclusively in the far-infrared motion detectors.

Two common types of Fresnel lenses are presently manufactured. One is a constant-step lens (Fig. 4.14A) and the other is a constant depth lens (Fig. 4.14B). In practice, it is difficult to maintain a curved surface of each small groove; hence, the profile of a groove is approximated by a flat surface. This demands that the steps be positioned close to each other. In fact, the closer the steps, the more accurate the lens.

In a constant-step lens, a slope angle φ of each groove is a function of its distance h from the optical axis. As a result, the depths of the grooves increase with the distance

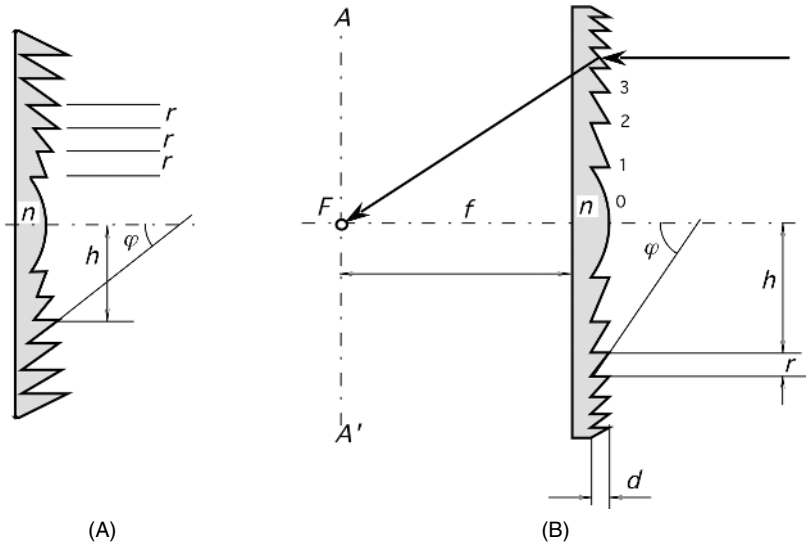


Fig. 4.14. Constant-step (A) and constant-depth (B) Fresnel lenses.

from the center. A central portion of the lens may be flat if its diameter is at least 20 times smaller than the focal length. For the shorter focal lengths, it is good practice to maintain a spherical profile of a central portion. The slope angle of each step may be determined from the following formula, which is valid for small values of h:

$$\varphi = \arctan\left(\frac{hn}{f(n-1)}\right), \tag{4.28}$$

where f is the focal length.

For a constant-depth lens, both the slope angle φ and the step distance r vary with the distance from the center. The following equations may be useful for the lens calculation. The distance of a groove from the center may be found through the groove number ξ (assuming the center portion has number 0);

$$h = \sqrt{2f(n-1)\xi\, d - \xi^2 d^2}, \tag{4.29}$$

and the slope angle is

$$\varphi = \arcsin\left(\frac{h}{(n-1)f}\right). \tag{4.30}$$

The total number of grooves in the lens may be found through a Fresnel lens aperture (maximum dimension) D:

$$\Gamma = \frac{(n-1)f - \sqrt{f^2(n-1)^2 - D^2}}{d}. \tag{4.31}$$

The Fresnel lens may be slightly bent if it is required for a sensor design. However, a bend changes the positions of the focal points. If a lens is bent with its groves inside the curvature, the focal distance decreases.

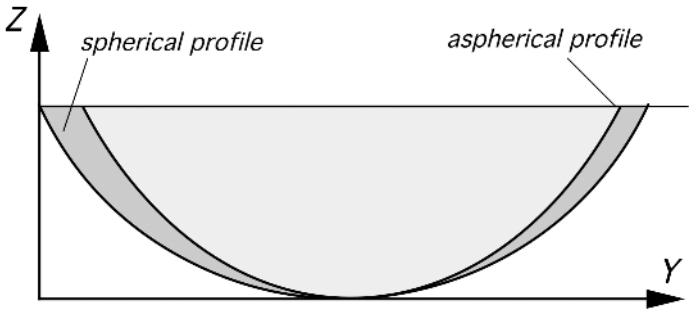


Fig. 4.15. Comparison of spherical and aspheric lens profiles.

It is known that a spherical surface of a lens will produce a spherical aberration. Therefore, for applications where high-quality focusing is required, the continuous surface from which the contours of the groves are determined should not be spherical, but aspherical. The profile of a continuous aspherical surface can often be described by a standard equation of a conic, axially symmetrical about the z axis (Fig. 4.15):

$$Z = \frac{CY^2}{1 + \sqrt{1 - (K+1)C^2Y^2}}, \tag{4.32}$$

where Z and Y are the coordinates of the surface, C is the vertex curvature, and K is the conic constant. The vertex curvature and the conic constant can be chosen depending on the desired characteristics of the lens, and the contours of each grove can be figured using this equation. C and K will depend on several factors, such as the desired focal length, the index of refraction, and the particular application.

4.7 Fiber Optics and Waveguides

Although light does not go around the corner, it can be channeled along complex paths by the use of waveguides. To operate in the visible and near-infrared spectral ranges, the guides may be fabricated of glass or polymer fibers. For the mid- and far-infrared spectral ranges, the waveguides are made as hollow tubes with highly reflective inner surfaces. The waveguide operates on the principle of the internal reflections where light beams travel in a zigzag pattern. A fiber can be used to transmit light energy in the otherwise inaccessible areas without any transport of heat from the light source. The surface and ends of a round or other cross-section fiber are polished. An outside cladding may be added. When glass is hot, the fibers can be bent to curvature radii of 20–50 times their section diameter and after cooling, to 200–300 diameters. Plastic fibers fabricated of polymethyl methacrylate may be bent at much smaller radii than glass fibers. A typical attenuation for a 0.25-mm polymer fiber is in the range of 0.5 dB/meter of length. Light propagates through a fiber by means of a total internal reflection, as shown in Fig. 4.16B. It follows from Eq. (4.23) that light passing to air

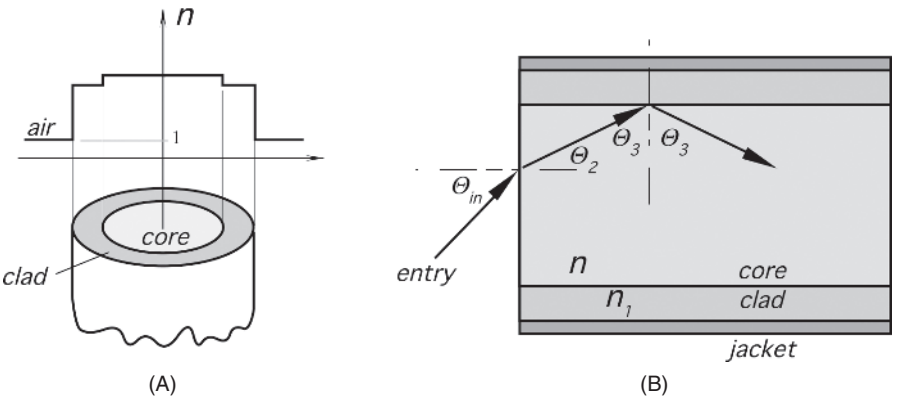


Fig. 4.16. Optical fibers: (A) a step-index multiple fiber; (B) determination of the maximum angle of entry.

from a medium having a refractive index n is subject to the limitation of an angle of total internal reflection. In a more general form, light may pass to another medium (cladding) having refractive index n_1; then, Eq. (4.23) becomes

$$\Theta_0 = \arcsin\left(\frac{n_1}{n}\right). \tag{4.33}$$

Figure 4.16A shows a profile of the index of refraction for a single fiber with the cladding where the cladding must have a lower index of refraction to assure a total internal reflection at the boundary. For example, a silica-clad fiber may have compositions set so that the core (fiber) material has an index of refraction of 1.5 and the clad has an index of refraction of 1.485. To protect the clad fiber, it is typically enclosed in some kind of protective rubber or plastic jacket. This type of the fiber is called a "step index multimode" fiber, which refers to the profile of the index of refraction.

When light enters the fiber, it is important to determine the maximum angle of entry which will result in total internal reflections (Fig. 4.16B). If we take that minimum angle of an internal reflection $\Theta_0 = \Theta_3$, then the maximum angle Θ_2 can be found from Snell's law:

$$\Theta_{2(\max)} = \arcsin\left(\frac{\sqrt{n^2 - n_1^2}}{n}\right). \tag{4.34}$$

Applying Snell's law again and remembering that for air $n \approx 1$, we arrive at

$$\sin\Theta_{\text{in}(\max)} = n_1 \sin\Theta_{2(\max)}. \tag{4.35}$$

Combining Eqs. (4.34) and (4.35), we obtain the largest angle with the normal to the fiber end for which the total internal reflection will occur in the core:

$$\Theta_{\text{in}(\max)} = \arcsin\left(\sqrt{n^2 - n_1^2}\right). \tag{4.36}$$

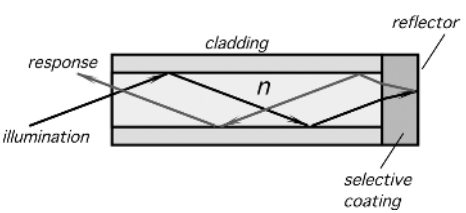
Light rays entering the fiber at angles greater than $\Theta_{in(max)}$ will pass through to the jacket and will be lost. For data transmission, this is an undesirable event, however, in a specially designed fiber-optic sensor, the maximum entry angle can be a useful phenomenon for modulating light intensity.

Sometimes, the value $\Theta_{in(max)}$ is called the numerical aperture of the fiber. Due to variations in the fiber properties, bends, and skewed paths, the light intensity does not drop to zero abruptly but rather gradually diminishes to zero while approaching $\Theta_{in(max)}$. In practice, the numerical aperture is defined as the angle at which light intensity drops by some arbitrary number, (e.g., −10 dB of the maximum value).

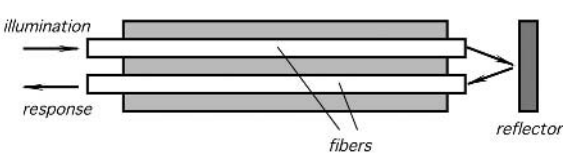
One of the useful properties of fiber-optic sensors is that they can be formed into a variety of geometrical shapes depending on the desired application. They are very useful for the design of miniature optical sensors which are responsive to such stimuli, as pressure, temperature, chemical concentration, and so forth. The basic idea for use of fiber optics in sensing is to modulate one or several characteristics of light in a fiber and, subsequently, to optically demodulate the information by conventional methods. A stimulus may act on a fiber either directly or it can be applied to a component attached to the fiber's outer surface or the polished end to produce an optically detectable signal.

To make a fiber chemical sensor, a special solid phase of a reagent may be formed in the optical path coupled to the fiber. The reagent interacts with the analyte to produce an optically detectable effect, (e.g., to modulate the index of refraction or coefficient of absorption). A cladding on a fiber may be created from a chemical substance whose refractive index may be changed in the presence of some fluids [3]. When the angle of total internal reflection changes, the light intensity varies.

Optical fibers may be used in two modes. In the first mode (Fig. 4.17A), the same fiber is used to transmit the excitation signal and to collect and conduct an optical response back to the processing device. In the second mode, two or more fibers are



(A)



(B)

Fig. 4.17. (A) Single- and (B) dual-fiber-optic sensors.

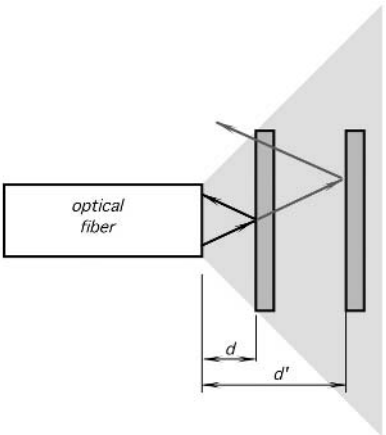


Fig. 4.18. Fiber-optic displacement sensor utilizes the modulation of reflected light intensity.

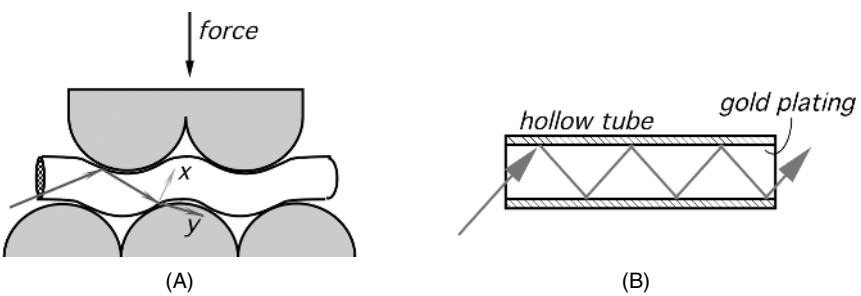


Fig. 4.19. Fiber-optic microbend strain gauge (A) and a waveguide for the far-infrared radiation (B).

employed where excitation (illumination) function and collection function are carried out by separate fibers (Fig. 4.17B).

The most commonly used type of fiber-optic sensor is an intensity sensor, where light intensity is modulated by an external stimulus [4]. Figure 4.18 shows a displacement sensor where a single-fiber waveguide emits light toward the reflective surface. Light travels along the fiber and exits in a conical profile toward the reflector. If the reflector is close to the fiber end (distance d), most of the light is reflected into the fiber and propagates back to the light detector at the other end of the fiber. If the reflector moves away, some of the rays are reflected outside of the fiber end, and fewer and fewer photons are returned back. Due to the conical profile of the emitted light, a quasilinear relationship between the distance d and the intensity of the returned light can be achieved over a limited range.

The so-called microbend strain gauge can be designed with an optical fiber which is squeezed between two deformers, as shown in Fig. 4.19A. The external force applied to the upper deformer bends the fiber affecting a position of an internal reflective surface. Thus, a light beam which normally would be reflected in direction

x approaches the lower part of the fiber at an angle which is less than Θ_0—the angle of total internal reflection [Eq. (4.33)]. Thus, instead of being reflected, light is refracted and moves in the direction y through the fiber wall. The closer the deformers come to each other, the more light goes astray and the less light is transmitted along the fiber.

For operation in the spectral range where loss in fibers is too great, hollow tubes are generally used for light channeling (Fig. 4.19B). The tubes are highly polished inside and coated with reflective metals. For instance, to channel thermal radiation, a tube may be fabricated of brass and coated inside by two layers: nickel as an underlayer to level the surface, and the optical-quality gold having thickness in the range 500–1000Å. Hollow waveguides may be bent to radii of 20 or more of their diameters. Although fiber optics use the effect of the total internal reflection, tubular waveguides use a first surface mirror reflection, which is always less than 100%. As a result, loss in a hollow waveguide is a function of a number of reflections; that is, loss is higher for the smaller diameter and the longer length of a tube. At length/diameter ratios more than 20, hollow waveguides become quite inefficient.

4.8 Concentrators

Regarding optical sensors and their applications, there is an important issue of the increasing density of the photon flux impinging on the sensor's surface. In many cases, when only the energy factors are of importance and a focusing or imaging is not required, special optical devices can be used quite effectively. These are the so-called nonimaging collectors, or concentrators [5]. They have some properties of the waveguides and some properties of the imaging optics (like lenses and curved mirrors). The most important characteristic of a concentrator is the ratio of the area of the input aperture divided by the area of the output aperture. The concentration ratio C is always more than unity; that is, the concentrator collects light from a larger area and directs it to a smaller area (Fig. 4.20A) where the sensing element is positioned. There is a theoretical maximum for C:

$$C_{\max} = \frac{1}{\sin^2 \Theta_i}, \tag{4.37}$$

where Θ_i is the maximum input semiangle. Under these conditions, the light rays emerge at all angles up to $\pi/2$ from the normal to the exit face. This means that the exit aperture diameter is smaller by $\sin \Theta_i$ times the input aperture. This gives an advantage in the sensor design, as its linear dimensions can be reduced by that number while maintaining a near equal efficiency. The input rays entering at angle Θ will emerge within the output cone with angles dependent of point of entry.

The concentrators can be fabricated with reflective surfaces (mirrors) or refractive bodies, or as a combination of both. A practical shape of the reflective parabolic concentrator is shown in Fig. 4.20B. It is interesting to note that cone light receptors in the retina of a human eye have a shape similar to that shown in Fig. 4.20B [6].

The tilted parabolic concentrators have very high efficiency[2]: They can collect and concentrate well over 90% of the incoming radiation. If a lesser efficiency is ac-

[2] This assumes that the reflectivity of the inner surface of the concentrator is ideal.

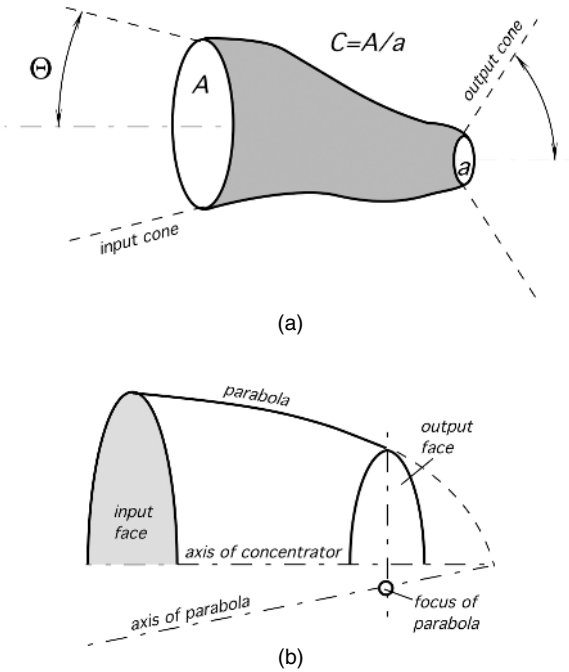


Fig. 4.20. Nonimaging concentrator (A) a general schematic; (B) a concentrator having a parabolic profile.

ceptable, a conical rather than paraboloid concentrator can be employed. Some of the incoming rays will be turned back after several reflections inside the cone; however, its overall efficiency is still near 80%. Clearly, the cones are easier to fabricate than the paraboloids of revolution.

4.9 Coatings for Thermal Absorption

All of thermal radiation sensors, either Passive Infrared (PIR) or Active Infrared (AFIR), rely on the absorption or emission of electromagnetic waves in the far-infrared spectral range. According to Kirchhoff's discovery, absorptivity α and emissivity ε are the same thing (see Section 3.12.3 of Chapter 3). Their value for the efficient sensor's operation must be maximized, that is, it should be made as close to unity as possible. This can be achieved by either a processing surface of a sensor to make it highly emissive, or by covering it with a coating having a high emissivity. Any such coating should have a good thermal conductivity and a very small thermal capacity, which means that it must be very thin.

Several methods are known to give a surface the emissive properties. Among them are a deposition of thin metal films (like nichrome) having reasonably good emissivity, a galvanic deposition of porous platinum black [7], and evaporation of

metal in atmosphere of low-pressure nitrogen [8]. The most effective way of creating a highly absorptive (emissive) material is to form it with a porous surface [9]. Particles with sizes much smaller than the wavelength generally absorb and diffract light. The high emissivity of a porous surface covers a broad spectral range; however, it decreases with the increased wavelength. A film of gold black with a thickness corresponding to 500 $\mu g/cm^2$ has an emissivity of over 0.99 in the near-, mid- and far-infrared spectral ranges.

To form porous platinum black, the following electroplating recipe can be used [10]:

Platinum chloride H_2PTCl_6 aq:	2 g
Lead acetate $Pb(OOCCH_3)_2 \cdot 3H_2O$:	16 mg
Water:	58 g

Out of this galvanic bath, the films were grown at room temperature on silicon wafers with a gold underlayer film. A current density was 30 mA/cm^2. To achieve an absorption better than 0.95, a 1.5 g/cm^2 film is needed.

To form a gold black by evaporation, the process is conducted in a thermal evaporation reactor at a nitrogen atmosphere of 100 Pa pressure. The gas is injected via a microvalve, and the gold source is evaporated from the electrically heated tungsten wire from a distance of about 6 cm. Due to collisions of evaporated gold with nitrogen, the gold atoms lose their kinetic energy and are slowed down to thermal speed. When they reach the surface, their energy is too low to allow surface mobility and they stick to the surface on the first touch event. Gold atoms form a surface structure in the form of needles with linear dimensions of about 25 nm. The structure resembles surgical cotton wool. For the best results, gold black should have a thickness in the range from 250 to 500 $\mu g/cm^2$.

Another popular method to enhance emissivity is to oxidize a surface metal film to form metal oxide, which, generally, is highly emissive. This can be done by metal deposition in a partial vacuum.

Another method of improving the surface emissivity is to coat a surface with an organic paint (visible color of the paint is not important). These paints have far-infrared emissivity from 0.92 to 0.97; however, the organic materials have low thermal conductivity and cannot be effectively deposited with thicknesses less than 10 μm. This may significantly slow the sensor's speed response. In micromachined sensors, the top surface may be given a passivation glass layer, which not only provides an environmental protection but has an emissivity of about 0.95 in the far-infrared spectral range.

4.10 Electro-optic and Acousto-optic Modulators

One of the essential steps of a stimulus conversion in optical sensors is their ability to modify light in some way (e.g., to alter its intensity by a control signal). This is called modulation of light. The control signal can have different origins: temperature, chemical compounds with different refractive indices, electric filed, mechanical stress, and so forth. Here, we examine light modulation by electric signals and acoustic waves.

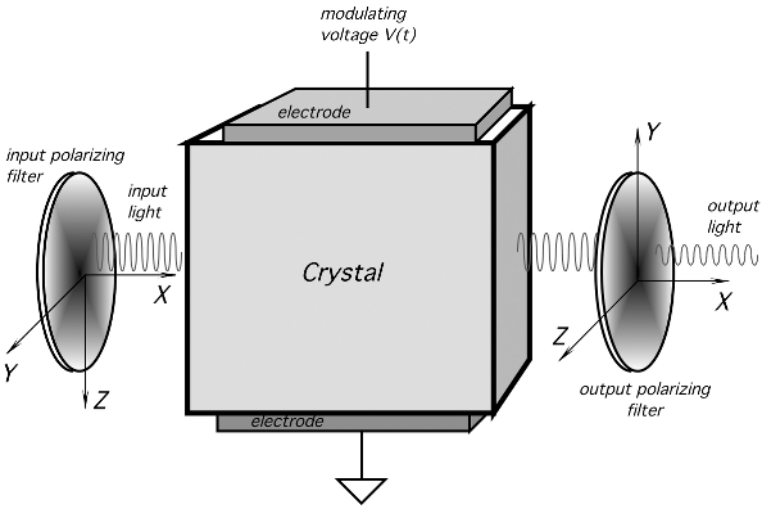


Fig. 4.21. Electro-optic modulator consists of two polarizing filters and a crystal.

In some crystals, the refractive index can be linked to an applied electric field [11]. The effect is characterized in the context of the propagation of a light beam through a crystal. For an arbitrary propagation direction, light maintains constant linear polarization through a crystal for only those polarization directions allowed by the crystal symmetry. An external electric field applied to a crystal may change that symmetry, thus modulating the light intensity. Lithium niobate ($LiNbO_3$) is one of the most widely used materials for electro-optic devices. A crystal is positioned between two polarizing filters which are oriented at 90° with respect to one another (Fig. 4.21). The input polarizer is oriented at 45° to the axis of the modulating crystal [12]. The crystal modulator has two electrodes attached to its surface. By changing the modulator voltage, the polarization of the light incident on the output polarizer is varied, which, in turn. leads to the intensity modulation.

A similar effect can be observed when the crystal is subjected to mechanical effects—specifically, to acoustic waves [11,13]. However, acousto-optic devices are used most often in fiber-optic applications as optical frequency shifters, and only to a lesser extent as intensity modulators. In the modulator, the light beam propagating through a crystal interacts with a traveling-wave index perturbation generated by an acoustic wave. The perturbation results from a photoelastic effect, whereby a mechanical strain produces a linear variation in refractive index. This resembles a traveling-wave diffraction grating, which, under certain conditions, can effectively deflect an optical beam (Fig. 4.22). Acousto-optic devices are often fabricated from lithium niobate and quartz, because acoustic waves can effectively propagate through these crystal over a frequency range from tens of megahertz to several gigahertz. The acoustic velocity in lithium niobate is about 6×10^3 m/s; thus a 1-GHz acoustic wave has a wavelength of about 6 μm, which is comparable to light in the infrared spectral range.

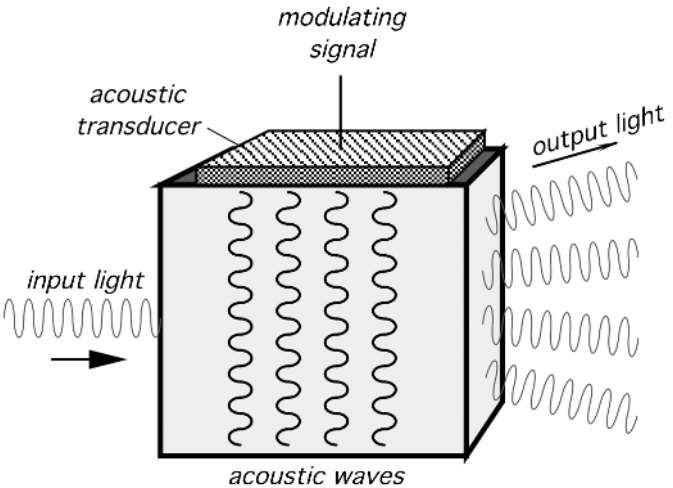


Fig. 4.22. Acousto-optic modulator produces multiple diffracted beams.

4.11 Interferometric Fiber-optic Modulation

Light intensity in an optical fiber can be modulated to produce a useful output signal. Figure 4.23 illustrates an optical waveguide which is split into two channels [13]. The waveguides are formed in a $LiNbO_3$ substrate doped with Ti to increase its refractive index. They are fabricated by a standard photolithographic lift-off technique. The substrate is patterned using a photomask. The Ti layer is electron-beam evaporated over the material; then, the photoresist is removed by a solvent, leaving the Ti in the waveguide pattern. The Ti atoms later diffuse into the substrate by baking [11]. This process results in a graded refractive index profile with a maximum difference at the surface of about 0.1% higher than the bulk value. Light is coupled to the waveguides through polished end faces. Electrodes are positioned in parallel to the waveguides, which recombine at the output. Voltage across the electrodes produces a significant phase shift in the light waves.

The optical transmission of the assembly varies sinusoidally with the phase shift $\Delta\phi$ between two recombined signals, which is controlled by voltage $V(t)$:

$$\frac{P_{\text{out}}}{P_{\text{in}}} = \frac{1}{2}\left\{1 + \cos\left[\frac{\pi V(t)}{V\pi} + \Theta_B\right]\right\}, \tag{4.38}$$

where $V\pi$ is the voltage change required for the full on–off modulation and Θ_B is the constant which can be adjusted for the optimum operating point. When the phase difference between the light in two waveguides before the recombination $\Delta\phi = 0$, the output couples to the output waveguide. When the shift $\Delta\phi = \pi$, light propagates into the substrate. The well-designed modulators can achieve a high extinction ratio on the order of 30 dB.

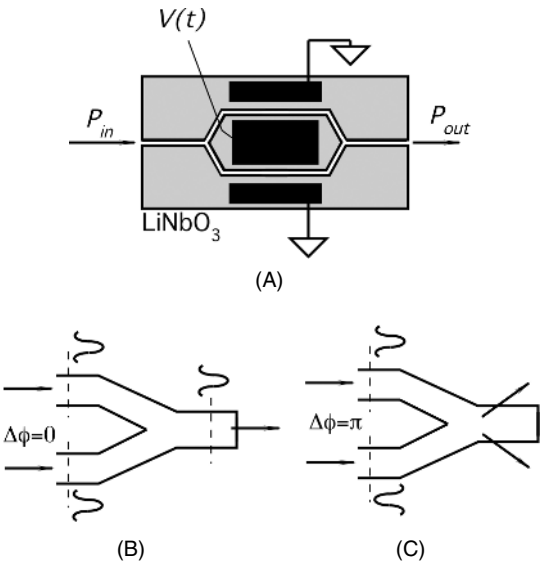


Fig. 4.23. Channel-waveguide interferometric intensity modulator (A). Light recombines in the exit fiber when the phase shift is zero (B). Light radiates into the substrate when the phase shift is π (C). (Adapted from Ref. [12].)

References

1. Begunov, B. N., Zakaznov, N. P., Kiryushin, S. I.,and Kuzichev, V. I. *Optical Instrumentation. Theory and Design.* Mir Publishers, Moscow, 1988.
2. *Applications of Phototransistors in Electro-optic Systems*. Motorola, 1988.
3. Giuliani, J. F. Optical waveguide chemical sensors. In: *Chemical Sensors and Microinstrumentation*. Murray, R.W. et al (eds) American Chemical Society, Washington, DC, 1989, Chap. 24.
4. Mitchell G. L. Intensity-based and Fabry–Perot interferometer sensors. In: *Fiber Optic Sensors: An Introduction for Engineers and Scientists*. E. Udd, ed., John Wiley & Sons, New York, 1991, Chap. 6.
5. Welford, W. T., and Winston, R. *High Collection Nonimaging Optics*. Academic Press, San Diego, CA, 1989.
6. Winston, R., and Enoch, J.M. Retinal cone receptor as an ideal light collector. *J. Opt. Soc. Am.* 61, 1120–1121, 1971.
7. von Hevisy, G. and Somiya, T. Über platinschwarz. *Zeitschr. Phys. Chemie A* 171, 41, 1934.
8. Harris, L., McGinnes, R., and Siegel, B. *J. Opt. Soc. Am.* 38, 7, 1948.
9. Persky, M.J. Review of black surfaces for space-borne infrared systems. *Rev. Sci. Instrum.* 70(5) 2193–2217, 1999.

10. Lang, W., Kühl, K., and Sandmaier, H. Absorption layers for thermal infrared detectors. In: *Transducers'91. International Conference on Solid-state Sensors and Actuators. Digest of Technical Papers*. IEEE, New York, 1991, pp: 635–638.
11. Yariv, A. Optical electronics, 3rd ed. Holt, Reinhart and Winston, New York, 1985.
12. Johnson, L. M. Optical modulators for fiber-optic sensors. In: *Fiber Optic Sensors: Introduction for Engineers and Scientists*. E. Udd, ed. John Wiley & Sons, New York 1991.
13. Haus, H. A. *Waves and Fields in Optoelectronics*. Prentice-Hall, Englewood Cliffs, NJ, 1984.

4

Optical Components of Sensors

"Where the telescope ends, the microscope begins.
Which of the two has the grander view?"

—Victor Hugo

Light phenomena, such as reflection, refraction, absorption, interference, polarization, and speed, are the powerful utensils in a sensor designer's toolbox. Optical components help to manipulate light in many ways. In this chapter, we discuss these components from a standpoint of geometrical optics. When using geometrical optics, we omit properties of light which are better described by quantum mechanics and quantum electrodynamics. We will ignore not only quantum properties of light but the wave properties as well. We consider light as a moving front or a ray which is perpendicular (normal) to that front. To do so, we should not discuss any optical elements whose dimensions are too small compared to the wavelength. For example, if a glass window is impregnated with small particles of submicron size, we should completely ignore them for any geometrical calculations from the near-infrared to longer wavelengths. Another example is a diffractive grating. Its operation cannot be described by methods of geometrical optics. In such cases, the methods of quantum electrodynamics (QED) need to be used. Here, we summarize those optical elements most applicable for the sensor design. For more detailed discussions of geometrical optics, we refer the reader to special texts, (e.g., Ref. [1]).

Before light can be manipulated, first we need to have light generated. There are several ways to produce light. Some sources of light are natural and exist without our will or effort; some must be incorporated into a measurement device. The natural sources of light include celestial objects, such as the Sun, Moon, stars, and so forth. Also, natural sources of light include all material objects that radiate thermal energy depending on their temperatures, as it was covered in Chapter 3. These include fire, exothermic chemical reactions, living organisms, and other natural sources whose temperatures are different from their surroundings and whose thermal radiation can be selectively detected by the optical devices. The man-made sources of light include

5

Interface Electronic Circuits

5.1 Input Characteristics of Interface Circuits

A system designer is rarely able to connect a sensor directly to processing, monitoring, or recording instruments, unless a sensor has a built-in electronic circuit with an appropriate output format. When a sensor generates an electric signal, that signal often is either too weak or too noisy, or it contains undesirable components. In addition, the sensor output may be not compatible with the input requirements of a data acquisition system, that is, it may have a wrong format. To mate a sensor and a processing device, they either must share a "common value" or some kind of a "mating" device is required in between. In other words, the signal from a sensor usually has to be *conditioned* before it is fed into a processing device (a load). Such a load usually requires either voltage or current as its input signal. An interface or a *signal conditioning* circuit has a specific purpose: to bring the signal from the sensor up to the format which is compatible with the load device. Figure 5.1 shows a stimulus that acts on a sensor which is connected to a load through an interface circuit. To do its job effectively, an interface circuit must be a faithful slave of two masters: the sensor and the load device. Its input characteristics must be matched to the output characteristics of the sensor and its output must be interfaceable with the load. This book, however, focuses on the sensors; therefore, we will discuss only the front stages of the interface circuits.

The input part of an interface circuit may be specified through several standard numbers. These numbers are useful for calculating how accurately the circuit can process the sensor's signal and what the circuit's contribution to a total error budget is.

The input impedance shows by how much the circuit loads the sensor. The impedance may be expressed in a complex form as

$$\boldsymbol{Z} = \frac{\boldsymbol{V}}{\boldsymbol{I}}, \tag{5.1}$$

where $\boldsymbol{V}$ and $\boldsymbol{I}$ are complex notations for the voltage and the current across the input impedance. For example, if the input of a circuit is modeled as a parallel connection of

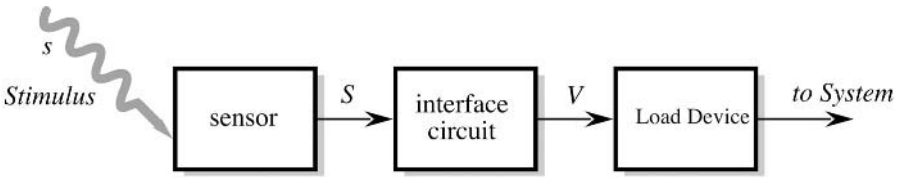


Fig. 5.1. Interface circuit matches the signal formats of a sensor and a load device.

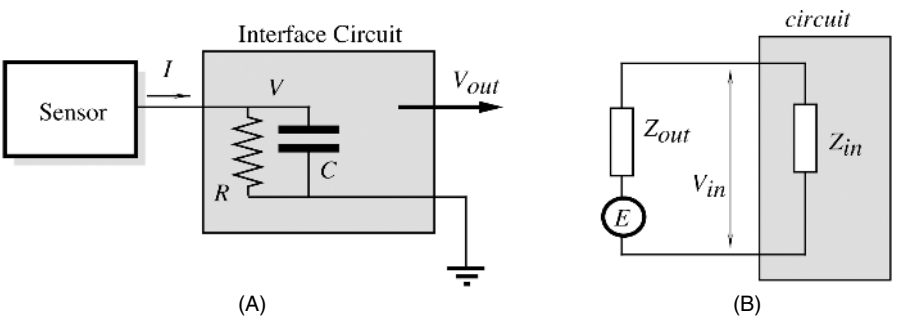


Fig. 5.2. Complex input impedance of an interface circuit (A) and equivalent circuit of a voltage-generating sensor (B).

input resistance R and input capacitance C (Fig. 5.2A), the complex input impedance may be represented as

$$\mathbf{Z} = \frac{R}{1 + j\omega RC}, \tag{5.2}$$

where ω is the circular frequency and $j = \sqrt{-1}$ is the imaginary unity. At very low frequencies, a circuit having a relatively low input capacitance and resistance has an input impedance almost equal to the input resistance: $Z \approx R$. Relatively low, here, means that the reactive part of the above equation becomes small; that is, the following holds

$$RC \ll \frac{1}{\omega}. \tag{5.3}$$

Whenever an input impedance of a circuit is considered, the output impedance of the sensor must be taken into account. For example, if the sensor is of a capacitive nature, to define a frequency response of the input stage, sensor's capacitance must be connected in parallel with the circuit's input capacitance. Formula (5.2) suggests that the input impedance is a function of the signal frequency. With an increase in the signal rate of change, the input impedance decreases.

Figure 5.2B shows an equivalent circuit for a voltage-generating sensor. The circuit is comprised of the sensor output, $\mathbf{Z}_{\text{out}}$, and the circuit input, $\mathbf{Z}_{\text{in}}$, impedances. The output signal from the sensor is represented by a voltage source, E, which is connected in series with the output impedance. Instead of a voltage source, for some sensors it is more convenient to represent the output signal as outgoing from a current

source, which would be connected in parallel with the sensor output impedance. Both representations are equivalent to one another, so we will use voltage. Accounting for both impedances, the circuit input voltage, V_{in}, is represented as

$$V_{\text{in}} = E\frac{Z_{\text{in}}}{Z_{\text{in}} + Z_{\text{out}}}. \tag{5.4}$$

In any particular case, an equivalent circuit of a sensor should be defined. This helps to analyze the frequency response and the phase lag of the sensor–interface combination. For instance, a capacitive detector may be modeled as pure capacitance connected in parallel with the input impedance. Another example is a piezoelectric sensor which can be represented by a very high resistance (on the order of $10^{11}\Omega$) shunted by a capacitance (in the order of 10 pF).

To illustrate the importance of the input impedance characteristics, let us consider a purely resistive sensor connected to the input impedance as shown in Fig. 5.2A. The circuit's input voltage as a function of frequency, f, can be expressed by

$$V = \frac{E}{\sqrt{1 + (f/f_c)^2}}, \tag{5.5}$$

where $f_c = (2\pi RC)^{-1}$ is the corner frequency, (i.e., the frequency where the amplitude drops by 3 dB). If we assume that a 1% accuracy in the amplitude detection is required, then we can calculate the maximum stimulus frequency that can be processed by the circuit:

$$f_{\max} \approx 0.14 f_c, \tag{5.6}$$

or $f_c \approx 7 f_{\max}$; that is, the impedance must be selected in such a way as to assure a sufficiently high corner frequency. For example, if the stimulus' highest frequency is 100 Hz, the corner frequency must be selected to be at least at 700 Hz. In practice, f_c is selected even higher, because of the additional frequency limitations in the subsequent circuits.

One should not overlook a speed response of the front stage of the interface circuit. Operational amplifiers, which are the most often used building blocks of interface circuits, usually have limited frequency bandwidths. There are the so-called programmable operational amplifiers which allow the user to control (to program) the bias current and, therefore, the first stage frequency response. The higher the current, the faster would be the response.

Figure 5.3 is a more detailed equivalent circuit of the input properties of a passive electronic interface circuit[1], (e.g., an amplifier or an A/D converter). The circuit is characterized by the input impedance $\mathbf{Z}_{\text{in}}$ and several generators. They represent voltages and currents which are generated by the circuit itself. These signals are spurious and may pose substantial problems if not handled properly. All of these signals are temperature dependent.

The voltage e_0 is called the input *offset voltage*. If the input terminals of the circuit are shorted together, that voltage would simulate a presence of an input dc signal

[1] Here, the word *passive* means that the circuit does not generate any excitation signal.

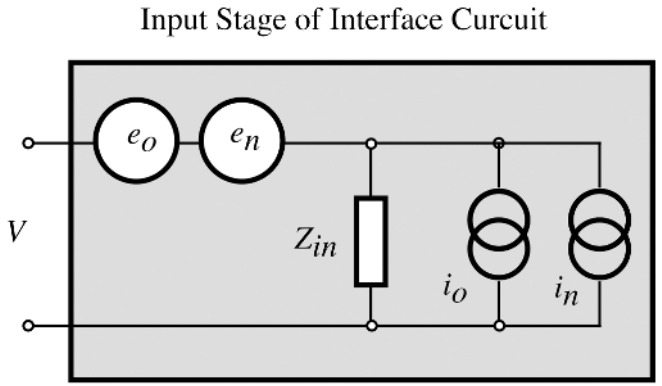


Fig. 5.3. Equivalent circuit of electrical noise sources at an input stage.

having a value of e_0. It should be noted that the offset voltage source is connected in series with the input and its resulting error is independent of the output impedance of the sensor.

The input *bias current* i_0 is also internally generated by the circuit. Its value is quite high for many bipolar transistors, much smaller for the JFETs, and even more lower for the CMOS circuits. This current may present a serious problem when a circuit or a sensor employs high-impedance components. The bias current passes through the input impedance of the circuit and the output impedance of the sensor, resulting in a spurious voltage drop. This voltage may be of a significant magnitude. For instance, if a piezoelectric sensor is connected to a circuit having an input resistance of 1 GΩ (10^9Ω) and the input bias current of 1 nA (10^{-9} A), the voltage drop at the input becomes equal to 1 GΩ × 1 nA=1 V—a very high value indeed. In contrast to the offset voltage, the bias current resulting error is proportional to the output impedance of the sensor. This error is negligibly small for the sensors having low output resistances. For instance, an inductive detector is not sensitive to the magnitude or variations in the bias current.

A circuit board *leakage current* may be a source of errors while working with high-impedance circuits. This current may be the result of lower surface resistance in the printed circuit board (PCB). Possible causes are poor-quality PCB material, surface contamination with solder flux residue (a poorly washed board), moisture, and degraded conformal coating. Figure 5.4A shows that a power-supply bus and the board resistance, R_L, may cause leakage current, i_L, through the sensor's output impedance. If the sensor is capacitive, its output capacitance will be charged very quickly by the leakage current. This will not only cause an error but may even lead to the sensor's destruction.

There are several techniques known to minimize the board leakage current effect. One is a careful board layout to keep higher-voltage conductors away from the high-impedance components. A leakage through the board thickness in multilayer boards should not be overlooked. Another method is electrical guarding, which is an old trick. The so-called driven shield is also highly effective. Here, the input circuit

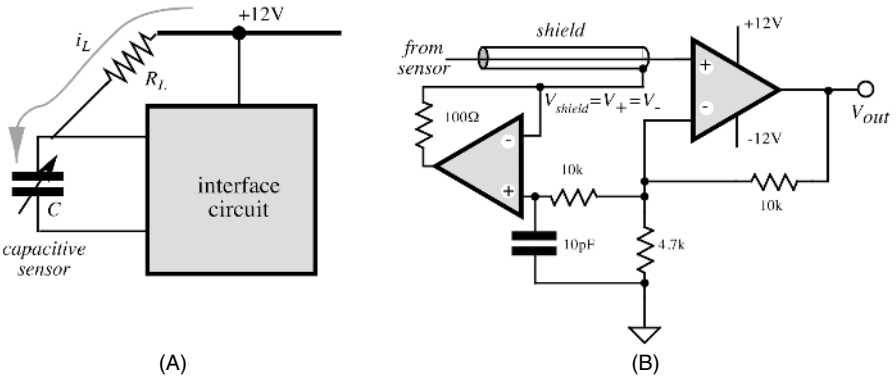


Fig. 5.4. Circuit board leakage affects input stage (A); driven shield of the input stage (B).

is surrounded by a conductive trace that is connected to a low-impedance point at the same potential as the input. The guard absorbs the leakage from other points on the board, drastically reducing currents that may reach the input terminal. To be completely effective, there should be guard rings on both sides of the printed circuit board. As an example, an amplifier is shown with a guard ring, driven by the relatively low impedance of the amplifier's inverting input.

It is highly advisable to locate the high-impedance interface circuits as close as possible to the sensors. However, sometimes connecting lines cannot be avoided. Coaxial shielded cables with good isolation are recommended [1]. Polyethylene or virgin (not reconstructed) Teflon is best for the critical applications. In addition to potential insulation problems, even short cable runs can reduce bandwidth unacceptably with high source resistances. These problems can be largely avoided by bootstrapping the cable's shield. Figure 5.4B shows a voltage follower connected to the inverting input of an amplifier. The follower drives the shield of the cable, thus reducing the cable capacitance, the leakage, and spurious voltages resulting from cable flexing. A small capacitance at the follower's noninverting input improves its stability.

Another problem that must be avoided is connecting to the input of an amplifier any components, besides a sensor, that potentially may cause problems. An example of such a "troublemaker" is a ceramic capacitor. In the hope of filtering out high-frequency transmitted noise at the input, a designer quite frequently uses filter capacitors either at the input, or in the feedback circuit of an input stage. If, for cost-saving or space-saving reasons, he selects a ceramic capacitor, he may get what he is not expecting. Many capacitors possess the so-called dielectric absorption properties which are manifested as a memory effect. If such a capacitor is subjected to a charge spike either from a sensor or from a power supply, or just from any external noise source, the charge will alter the capacitor's dielectric properties in such a way that the capacitor now behaves like a small battery. That "battery" may take a long tome to lose its charge—from few seconds to many hours. The voltage generated by that "battery" is added to the sensor's signal and may cause significant errors. If a capacitor must be employed at the input stage, a film capacitor should be used instead of ceramic.

5.2 Amplifiers

Most passive sensors produce weak output signals. The magnitudes of these signals may be on the order of microvolts (μV) or picoamperes (pA). On the other hand, standard electronic data processors, such as A/D converters, frequency modulators, data recorders, and so forth, require input signals of sizable magnitudes—on the order of volts (V) and milliamperes (mA). Therefore, an amplification of the sensor output signals has to be made with a voltage gain up to 10,000 and a current gain up to 1,000,000. Amplification is part of signal conditioning. There are several standard configurations of amplifiers which might be useful for amplifying signals from various sensors. These amplifiers may be built of discrete components, such as semiconductors, resistors, capacitors, and inductors. Alternatively, the amplifiers are frequently composed of standard building blocks, such as operational amplifiers and various discrete components.

It should be clearly understood that the purpose of an amplifier is much broader than just increasing the signal magnitude. An amplifier may be also an impedance-matching device, an enhancer of a signal-to-noise ratio, a filter, and an isolator between input and output.

5.2.1 Operational Amplifiers

One of the principal building blocks for the amplifiers is the so-called *operational amplifier* or OPAM, which is either an integrated (monolithic) or hybrid (a combination of monolithic and discrete parts) circuit. An integrated OPAM may contain hundreds of transistors, as well as resistors and capacitors. An analog circuit designer, by arranging discrete components around the OPAM (resistors, capacitors, inductors, etc.), may create an infinite number of useful circuits—not only the amplifiers, but many others circuits as well. Operational amplifiers are also used as cells in custom-made integrated circuits of the analog or mixed-technology types. These circuits are called *application-specific integrated circuits* or ASICs for short. Below, we will describe some typical circuits with OPAM which are often used in conjunction with various sensors.

As a building block, a good operational amplifier has the following properties (a schematic representation of OPAM is shown in Fig. 5.5):

- Two inputs: one inverting (−) and the other is noninverting (+)
- A high input resistance (on the order of hundreds of MΩ or even GΩ)
- A low output resistance (a fraction of Ω)
- The ability to drive capacitive loads
- A low input offset voltage e_0 (few mV or even μV)
- A low input bias current i_0 (few pA or even less)
- A very high *open-loop gain* A_{OL} (at least 10^4 and preferably over 10^6); that is, the OPAM must be able to magnify (amplify) a voltage difference V_{in} between its two inputs by a factor of A_{OL}
- A high common-mode rejection ratio (CMRR); that is, the amplifier suppresses the in-phase equal magnitude input signals (common-mode signals) V_{CM} applied to both inputs

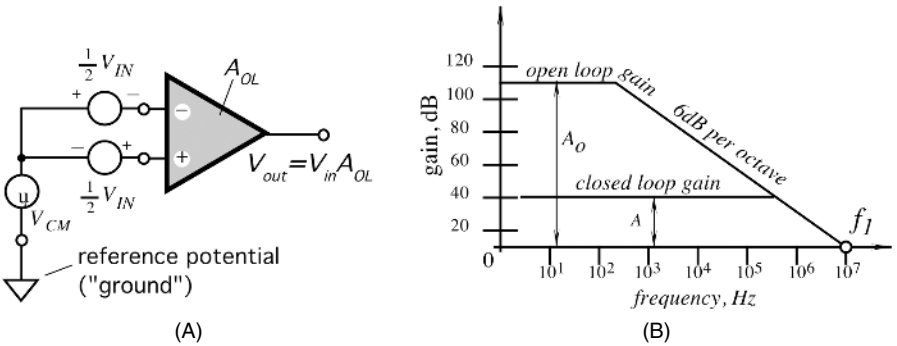


Fig. 5.5. General symbol of an operational amplifier (A) and gain/frequency characteristic of an OPAM (B).

- Low intrinsic noise
- A broad operating frequency range
- A low sensitivity to variations in the power-supply voltage
- A high environmental stability of its own characteristics

For detailed information and application guidance, the user should refer to data books published by the respective manufacturers. Such books usually contain selection guides for every important feature of an OPAM. For instance, OPAMs are grouped by such criteria as low offset voltages, low bias currents, low noise, bandwidth, and so forth.

Figure 5.5A depicts an operational amplifier without any feedback components. Therefore, it operates under the so-called *open-loop* conditions. An open loop gain, A_{OL}, of an OPAM is always specified but is not a very stable parameter. Its frequency dependence may be approximated by the graph in Fig. 5.5B. The A_{OL} changes with the load resistance, temperature, and the power-supply fluctuations. Many amplifiers have an open-loop gain temperature coefficient on the order of 0.2–1%/°C and a power-supply gain sensitivity on the order of 1%/%. An OPAM is very rarely used with an open loop (without the feedback components) because the high open-loop gain may result in circuit instability, a strong temperature drift, noise, and so forth. For instance, if the open-loop gain is 10^5, the input voltage drift of 10 μV would cause output drifts by about 1V.

The ability of an OPAM to amplify small-magnitude high-frequency signals is specified by the *gain–bandwidth product* (GBW) which is equal to the frequency f_1 where the amplifier gain becomes equal to unity. In other words, above the f_1 frequency, the amplifier cannot amplify. Figure 5.6A depicts a noninverting amplifier where resistors R_1 and R_2 define the feedback loop. The resulting gain, $A = 1 + R_2/R_1$, is a closed-loop gain. It may be considered constant over a much broader frequency range (see Fig. 5.5B); however, f_1 is the frequency-limiting factor regardless of the feedback. The linearity, gain stability, the output impedance, and gain accuracy are all improved by the amount of feedback and now depend mainly on characteristics of the feedback components. As a general rule for moderate accu-

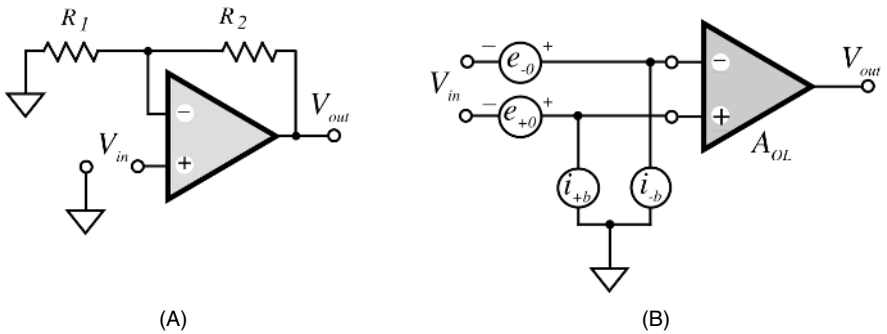


Fig. 5.6. Noninverting amplifier (A); offset voltages and bias currents in an operational amplifier represented by generators connected to its inputs (B).

racy, the open-loop gain of an OPAM should be at least 100 times greater than the closed-loop gain at the highest frequency of interest. For even higher accuracy, the ratio of the open-loop and closed-loop gains should be 1000 or more.

A typical datasheet for an OPAM specifies the bias and offset voltages. Due to limitations in manufacturing technologies, any OPAM acts not only as a pure amplifier but also as a generator of voltages and currents which may be related to its input (Fig. 5.6B). Because these spurious signals are virtually applied to the input terminals, they are amplified along with the useful signals.

Because of offset voltages and bias currents, an interface circuit does not produce zero output when a zero input signal is applied. In dc-coupled circuits, these undesirable input signals may be indistinguishable from the useful signal. If the input offset voltage is still too large for the desired accuracy, it can be trimmed out either directly at the amplifier (if the amplifier has dedicated trimming terminals) or in the independent offset compensation circuit.

An application engineer should be concerned with the output offset voltage, which can be derived from:

$$V_0 = A(e_0 + i_0 R_{\text{eqv}}) \tag{5.7}$$

where R_{eqv} is the equivalent resistance at the input (a combination of the sensor's output resistance and the input resistance of the amplifier), e_0 is the input offset voltage, and i_0 is the input bias current. The offset is temperature dependent. In circuits where the amplifier has high gain, the output voltage offset may be a source of substantial error. There are several ways to handle this difficulty. Among them is selecting an amplifier with low bias current, high input resistance, and low offset voltage. Chopper-stabilized amplifiers are especially efficient for reduction of offset voltages.

5.2.2 Voltage Follower

A voltage follower (Fig. 5.7) is a an electronic circuit that provides impedance conversion from a high level to a low level. A typical follower has high input impedance

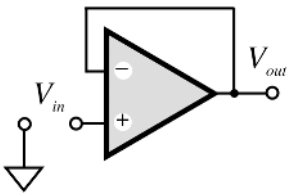


Fig. 5.7. Voltage follower with an operational amplifier.

(the high input resistance and the low input capacitance) and low output resistance (the output capacitance makes no difference). A good follower has a voltage gain very close to unity (typically, 0.999 at lower frequencies) and a high current gain. In essence, it is a current amplifier and impedance converter. Its high input and low output impedances make it indispensable for interfacing between many sensors and signal processing devices.

A follower, when connected to a sensor, has very little effect on the latter's performance, thus providing a buffering function between the sensor and the load. When designing a follower, these tips might be useful:

- For the current-generating sensors, the input bias current of the follower must be at least 100 times smaller than the sensor's current.
- The input offset voltage must be either trimable or smaller than the required least significant bit (LSB).
- The temperature coefficient of the bias current and the offset voltage should not result in errors of more than 1 LSB over an entire temperature range.

5.2.3 Instrumentation Amplifier

An instrumentation amplifier (IA) has two inputs and one output. It is distinguished from an operational amplifier by its finite gain (which is usually no more than 100) and the availability of both inputs for connecting to the signal sources. The latter feature means that all necessary feedback components are connected to other parts of the instrumentation amplifier, rather than to its noninverting and inverting inputs. The main function of the IA is to produce an output signal which is proportional to the difference in voltages between its two inputs:

$$V_{\text{out}} = A(V_+ - V_-) = A\Delta V, \tag{5.8}$$

where V_+ and V_- are the input voltages at noninverting and inverting inputs, respectively, and A is the gain. An instrumentation amplifier can be either built from an OPAM, in a monolithic or hybrid form. It is important to assure high input resistances for both inputs, so that the amplifier can be used in a true differential form. A differential input of the amplifier is very important for rejection of common-mode interferences having an additive nature (see Section 5.9). An example of a high-quality monolithic instrumentation amplifier is INA118 from Burr-Brown/Texas Instruments (www.ti.com). It offers a low offset voltage of 50 μV and a high common-mode rejection ratio (110 dB). The gain is programmed by a single resistor.

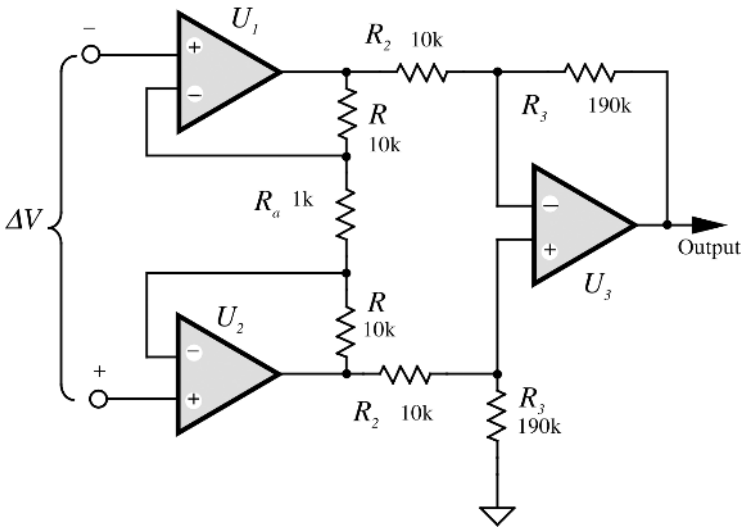


Fig. 5.8. Instrumentation amplifier with three operational amplifiers and matched resistors.

Although several monolithic instrumentation amplifiers are presently available, quite often discrete component circuits prove to be more cost-effective. A basic circuit of an IA is shown in Fig. 5.8. The voltage across R_a is forced to become equal to the input voltage difference ΔV. This sets the current through that resistor equal to $i = \Delta V / R_a$. The output voltages from the U_1 and U_2 OPAMs are equal to one another in amplitudes and opposite in the phases. Hence, the front stage (U_1 and U_2) has a differential input and a differential output configuration. The second stage (U_3) converts the differential output into a unipolar output and provides an additional gain. The overall gain of the IA is

$$A = \left(1 + \frac{2R}{R_a}\right) \frac{R_3}{R_2}. \tag{5.9}$$

The common-mode rejection ratio (CMRR) depends on matching of resistors within the corresponding group (R, R_2, and R_3). As a rule of thumb, 1% resistors yield CMRRs no better than 100, whereas for 0.1%, the CMRR is no better than 1000.

A good and cost-effective instrumentation amplifier can be built of two identical operational amplifiers and several precision resistors (Fig. 5.9A). The circuit uses the FET-input OPAMs to provide lower noise and lower input bias currents. U_1 acts as a noninverting amplifier and U_2 is the inverting one. Each input has a high impedance and can be directly interfaced with a sensor. A feedback from each amplifier forces voltage across the gain-setting resistor R_a to become equal to ΔV. The gain of the amplifier is equal to

$$A \approx 2\left(1 + \frac{R}{R_a}\right). \tag{5.10}$$

Hence, gain may vary from 2 (R_a is omitted) to a potentially open-loop gain ($R_a = 0$). With the components whose values are shown in Fig. 5.9A, the gain is $A = 100$. It

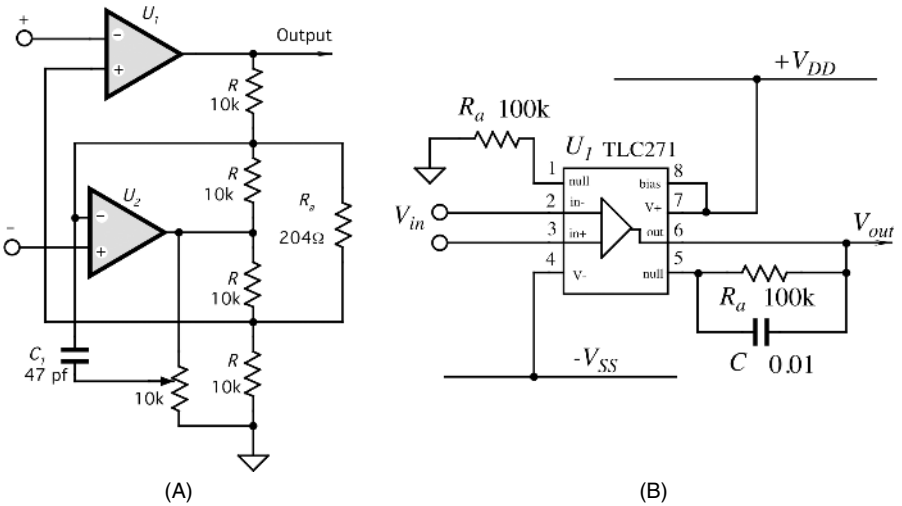


Fig. 5.9. (A) Instrumentation amplifier with two operational amplifiers; (B) low-cost instrumentation amplifier with one operational amplifier.

should be remembered, however, that the input offset voltage will be amplified with the same gain. The CMRR primarily depends on matching values of resistors R. At very low frequencies, it is the reciprocal of the net fractional resistor mismatch, [i.e., CMRR=10,000 (−80 dB) for a 0.01% mismatch]. At higher frequencies, the impedance mismatch must be considered, rather than the resistor mismatch. To balance the impedances, a trimpot and a capacitor C_1 may be used. Also, remember that IA as a rule requires a dual (plus and minus) power supply.

When cost is a really limiting factor and no high-quality dc characteristics are required, a very simple IA can be designed with just one operational amplifier and two resistors (Fig. 5.9B). The feedback resistor R_a is connected to the null-balance terminal of the OPAM, which is the output of the front stage of the monolithic circuit. The amount of the feedback through R_a depends on the actual circuit of an OPAM and somewhat varies from part to part. For the TLC271 operational amplifier (Texas Instruments), a gain of the circuit may be found from

$$A \approx 1 + \frac{R_a}{2\text{k}\Omega} \qquad (R_a \text{ is in k}\Omega), \tag{5.11}$$

which for values indicated in Fig. 5.9B gives a gain of about 50. The connections and values of the external components are different for different types of the operational amplifier. In addition, not all OPAMs can be used in such an unusual circuit.

5.2.4 Charge Amplifiers

The charge amplifier (CA) is a very special class of circuits which must have extremely low bias currents. These amplifiers are employed to convert to voltage signals from the capacitive sensors, quantum detectors, pyroelectric sensors, and other devices

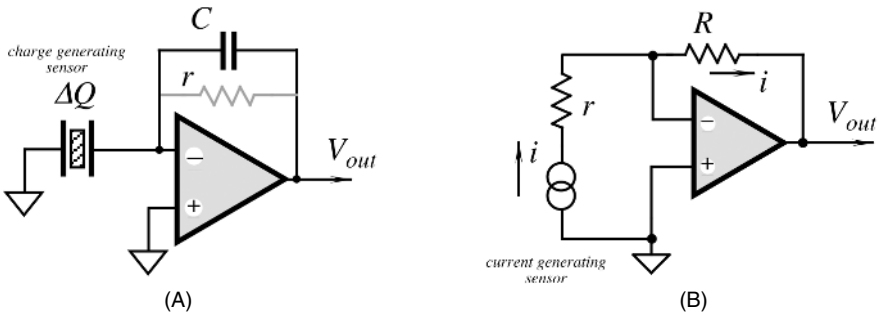


Fig. 5.10. Charge-to-voltage (A) and current-to-voltage (B) converters.

which generate very small charges (on the order of picocoulombs, pC) or currents (on the order of picoamperes). A basic circuit of a charge-to-voltage converter is shown in Fig. 5.10A. A capacitor, C, is connected into a feedback network of an OPAM. Its leakage resistance r must be substantially larger than the impedance of the capacitor at the lowest operating frequency. A good film capacitor is usually recommended along with a good quality printed circuit board where the components are coated with conformal coating.

A transfer function of the converter is

$$V_{\text{out}} = -\frac{\Delta Q}{C}. \tag{5.12}$$

Special hybrid charge-sensitive preamplifiers are available for precision applications. One example is DN630 from ThermOptics, Inc. (www.thermoptics.com). The amplifier can operate with sources of less than 1 pF capacitance. An internally connected 1-pF capacitor sets the gain of the amplifier to 1 V/pC sensitivity. The gain can be reduced by connecting one or a combination of the internal capacitor array to the input of the amplifier. It features low noise and has less than 5 ns rise and fall times.

Many sensors can be modeled as capacitors. Some capacitive sensors are active; that is, they require an excitation signal. Examples are microphones, capacitive force and pressure transducers, and humidity detectors. Other capacitive sensors are passive; that is, they directly convert a stimulus into an electric charge or current. Examples are the piezoelectric and pyroelectric detectors. There are also noncapacitive sensors that can be considered as current generators. An example is a photodiode.

A current-generating sensor is modeled by a leakage resistance, r, connected in parallel with a current generator that has an infinitely high internal resistance (Fig. 5.11). The sensor generates current, i, which has two ways to outflow: to the sensors leakage resistance, r, as current, i_0, and the other, i_{out}, toward the interface circuit input impedance, Z_L. Naturally, current i_0 is useless and to minimize the error of the current-to-voltage conversion, the leakage resistance of the sensor must be much larger than the impedance of the interface circuit.

Ohm's law suggests that to convert electric current i_{out} into voltage, current should pass through an appropriate impedance and the voltage drop across that impedance

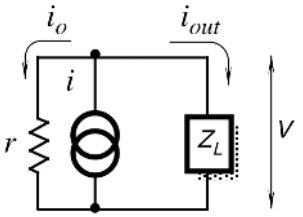


Fig. 5.11. An equivalent circuit of a current-generating sensor.

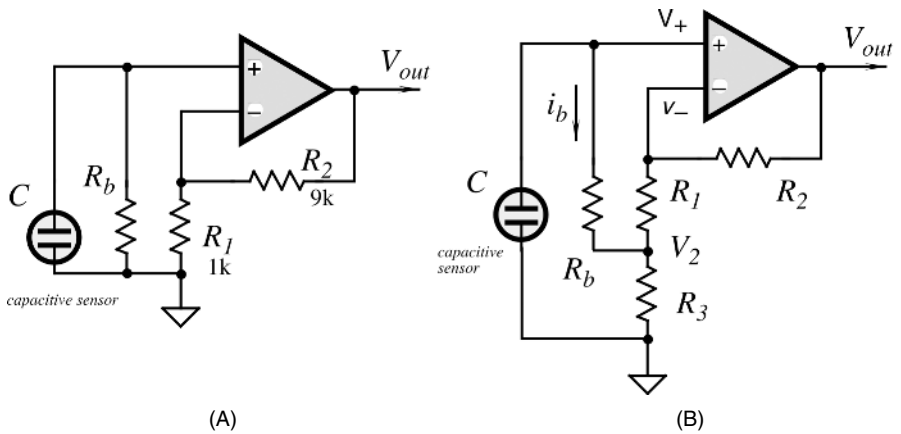


Fig. 5.12. Noninverting current-to-voltage converter (A) and resistance multiplier (B).

be proportional to the magnitude of the current. Fig. 5.10B shows a basic current-to-voltage converter where the current-generating sensor is connected to the inverting input of an OPAM, which serves as a virtual ground; that is, voltage at the inverting input is almost equal to that at the noninverting input, which is grounded. The sensor operates at nearly zero voltage across its terminals and its current is represented by the output voltage of the OPAM:

$$V_{\text{out}} = -i\,R. \tag{5.13}$$

A resistor, $r \ll R$ is often required for the circuit stability. At high frequencies, the OPAM would operate near the open-loop gain, which may result in oscillations. This is especially true when the sensor has reduced leakage resistance. The advantage of the virtual ground is that the output signal does not depend on the sensor's capacitance. The circuit produces a voltage whose phase is shifted by 180° with respect to the current. The noninverting circuit shown in Fig. 5.12A can convert and amplify the signal; however, its speed response depends on both the sensor's capacitance and the converting resistor R_1. Thus, the response to a step function in a time domain can be described by

$$V_{\text{out}} = i\,R_b \left(1 + \frac{R_2}{R_1}\right)(1 - e^{-t/rC}). \tag{5.14}$$

When converting currents from such sensors as piezoelectrics and pyroelectrics, the resistor R_b (R in Fig. 5.10B) may be required on the order of tens or even hundreds of gigohms. In many cases, resistors of such high values may be not available or impractical to use due to poor environmental stability. A high-ohmic resistor can be simulated by a circuit known as a resistance multiplier. It is implemented by adding a positive feedback around the amplifier. Figure 5.12B shows that R_1 and R_3 form a resistive divider. Due to a high open-loop gain of the OPAM, voltages at noninverting and inverting inputs are almost equal to one another: $V_+ \approx V_-$. As a result, voltage, V_2, at the divider is

$$V_2 = V_- \frac{R_3}{R_1 + R_3} \approx V_+ \frac{R_3}{R_1 + R_3}, \tag{5.15}$$

and current through the resistor is defined through the voltage drop:

$$I_b = \frac{\Delta V}{R_b} = \frac{V_+ - V_2}{R_b} = \frac{V_+}{R_b} \frac{R_1}{R_1 + R_3}. \tag{5.16}$$

From this equation, the input voltage can be found as a function of the input current and the resistive network:

$$V_+ = I_b R_b \left(1 + \frac{R_3}{R_1}\right). \tag{5.17}$$

It is seen that the resistor R_b is multiplied by a factor of $(1 + R_3/R_1)$. For example, if the highest resistor you may consider is 10 MΩ, by selecting the multiplication factor of, say, 5, you get a virtual resistance of 50 MΩ. Resistance multiplication, although being a powerful trick, should be used with caution. Specifically, noise, bias current, and offset voltage are all also multiplied by the same factor $(1 + R_3/R_1)$, which may be undesirable in some applications. Further, because the network forms a positive feedback, it may cause circuit instability. Therefore, in practical circuits, resistance multiplication should be limited to a factor of 10.

5.3 Excitation Circuits

External power is required for the operation of *active* sensors. Examples are temperature sensors [thermistors and resistive temperature detectors (RTDs)], pressure sensors (piezoresistive and capacitive), and displacement (electromagnetic and optical). The power may be delivered to a sensor in different forms. It can be a constant voltage, constant current, or sinusoidal or pulsing currents. It may even be delivered in the form of light or ionizing radiation. The name for that external power is an excitation signal. In many cases, stability and precision of the excitation signal directly relates to the sensor's accuracy and stability. Hence, it is imperative to generate the signal with such accuracy that the overall performance of the sensing system is not degraded. In the following subsections, we review several electronic circuits which feed sensors with appropriate excitation signals.

5.3.1 Current Generators

Current generators are often used as excitation circuits to feed sensors with predetermined currents that, within limits, are independent of the sensor properties, stimulus value, or environmental factors. In general terms, a current generator (current pump) is a device which produces electric current independent of the load impedance; that is, within the capabilities of the generator, the amplitude of its output current must remain substantially independent of any changes in the impedance of the load.

The usefulness of the current generators for the sensor interfaces is in their ability to produce excitation currents of precisely controlled magnitude and shape. Hence, a current generator should not only produce current which is load independent, but it also must be controllable from an external signal source (a waveform generator), which, in most cases, has a voltage output. A good current generator must produce current which follows the control signal with high fidelity and is independent of the load over a broad range of impedances.

There are two main characteristics of a current generator: the output resistance and the voltage compliance. The output resistance should be as high as practical. A voltage compliance is the highest voltage which can be developed across the load without affecting the output current. For a high resistive load, according to Ohm's law, a higher voltage is required for a given current. For instance, if the required excitation current is $i = 10$ mA and the highest load impedance at any given frequency is $Z_L = 10\,\text{k}\Omega$, a voltage compliance of at least $iZ_L = 100$ V would be needed. Below, we cover some useful circuits with increased voltage compliance where the output currents can be controlled by external signals.

A *unipolar* current generator is called either a current *source* (generates the outflowing current) or a current *sink* (generates the inflowing currents). Here, unipolar means that it can produce currents flowing in one direction only, usually toward the ground. Many of such generators utilize current-to-voltage characteristics of transistors. A voltage-controlled current source or sink may include an operational amplifier (Fig. 5.13A). In such a circuit, a precision and stable resistor R_1 defines the output current, i_{out}. The circuit contains a feedback loop through the OPAM that keeps voltage across resistor R_1 constant and, thus, the constant current. To deliver a higher current at a maximum voltage compliance, voltage drop as small as possible should be developed across the sensing resistor R_1. In effect, that current is equal to V_1/R_1. For better performance, the current through the base of the output transistor should be minimized; hence, a field-effect rather than bipolar transistor is often used as an output device.

It is well known that the transistor's collector current is very little dependent on collector voltages. This feature was employed by the so-called current mirrors. A current mirror has one current input and at least one (may be several) current output. Therefore, the output current is controlled by the input current. The input current is supplied from an external source and should be of a known value. Figure 5.13B shows the so-called Wilson current mirror, where voltage V_1 and resistance R_1 produce the input current i_{in}. The output transistor Q_1 acts as a current-controlled resistor, thus

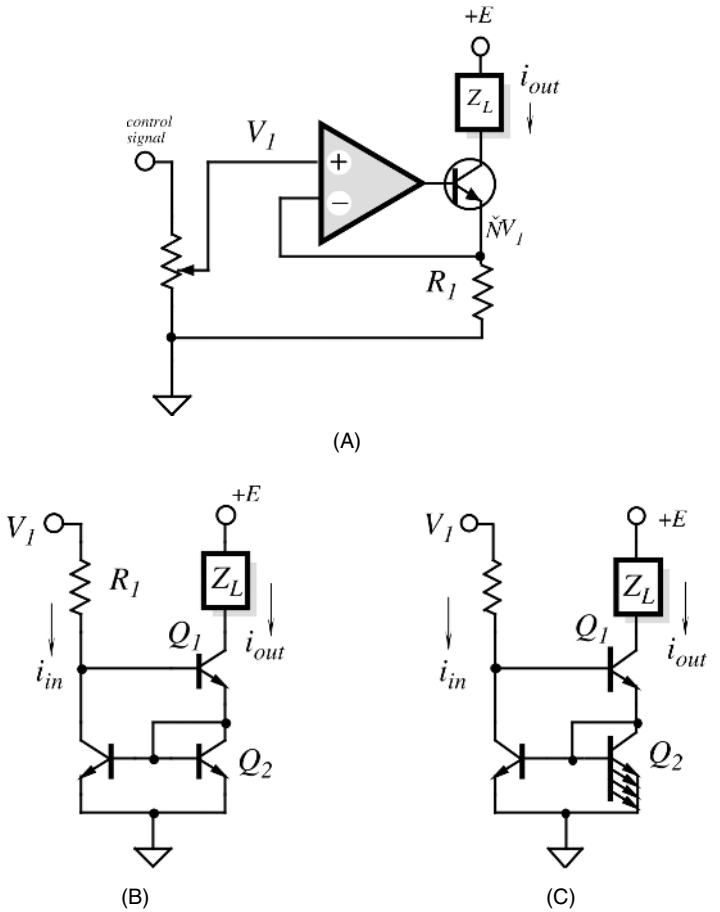


Fig. 5.13. Current sources: (A) with an OPAM, (B) current mirror; (C) current mirror with current multiplication.

regulating the output current i_{out} in such a manner as to maintain it equal to i_{in}. The output current may be multiplied several times if the transistor Q_2 (Fig. 5.13C) is fabricated with several emitters. Such a current sink is commercially available from Texas Instruments (part TLC014A). That current mirror has a voltage compliance of 35 V and an output resistance ranging from 2 to 200 MΩ (depending on the current).

For many applications, *bipolar* current generators may be required. Such a generator provides a sensor with an excitation current which may flow in both directions (inflowing and outflowing). Figure 5.14 shows noninverting and inverting circuits with an operational amplifier where the load is connected as a feedback. Current though the load Z_L, is equal to V_1/R_1, which is load independent. The load current follows V_1 within the operating limits of the amplifier. An obvious limitation of the circuit is that the load is "floating", that is, it is not connected to a ground bus or any

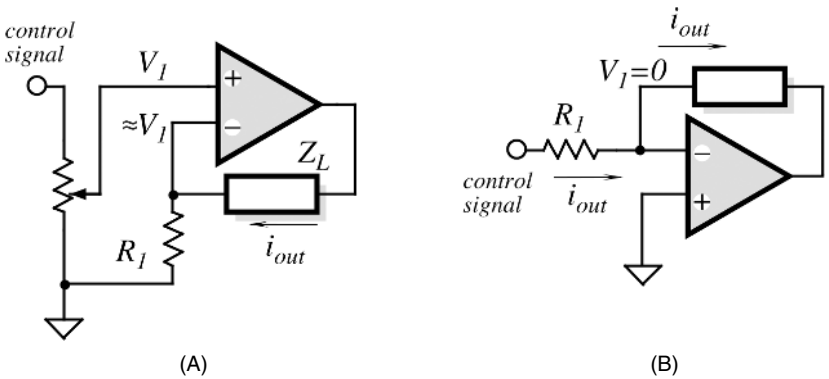


Fig. 5.14. Bipolar current generators with floating loads: (A) a noninverting circuit; (B) a circuit with a virtual ground.

other reference potential. For some applications, this is quite all right; however, many sensors need to be grounded or otherwise referenced. The circuit shown in Fig. 5.14B keeps one side of the load impedance near the ground potential, because a noninverting input of the OPAM is a virtual ground. Nevertheless, even in this circuit, the load is still fully isolated from the ground. One negative implication of this isolation may be an increased pickup of various kinds of transmitted noise.

In cases where the sensor must be grounded, a current pump invented by Brad Howland at MIT may be used (Fig. 5.15). The pump operation is based on utilizing both negative and positive feedbacks around the operational amplifier. The load is connected to the positive loop [2]. Current through the load is defined by

$$i_{\text{out}} = \frac{R_2}{R_1}\frac{(V_1 - V_2)}{R_5}. \tag{5.18}$$

A trimming resistor, P, must be adjusted to assure that

$$R_3 = R_1\frac{R_4 + R_5}{R_2}. \tag{5.19}$$

In that circuit, each resistor may have a relatively high value (100 kΩ or higher), but the value for R_5 should be relatively small. This condition improves the efficiency of the Howland current pump, as smaller voltage is wasted across R_5 and smaller current is wasted through R_4 and R_3. The circuit is stable for most of the resistive loads; however, to ensure stability, a few-picofarad capacitor C may be added in a negative feedback or/and from the positive input of an operational amplifier to ground. When the load is inductive, an infinitely large compliance voltage would be required to deliver the set current when a fast transient control signal is applied. Therefore, the current pump will produce a limited rising slope of the output current. The flowing current will generate an inductive spike across the output terminal, which may be fatal to the operational amplifier. For the large inductive load, it is advisable to clamp the load with diodes to the power-supply buses.

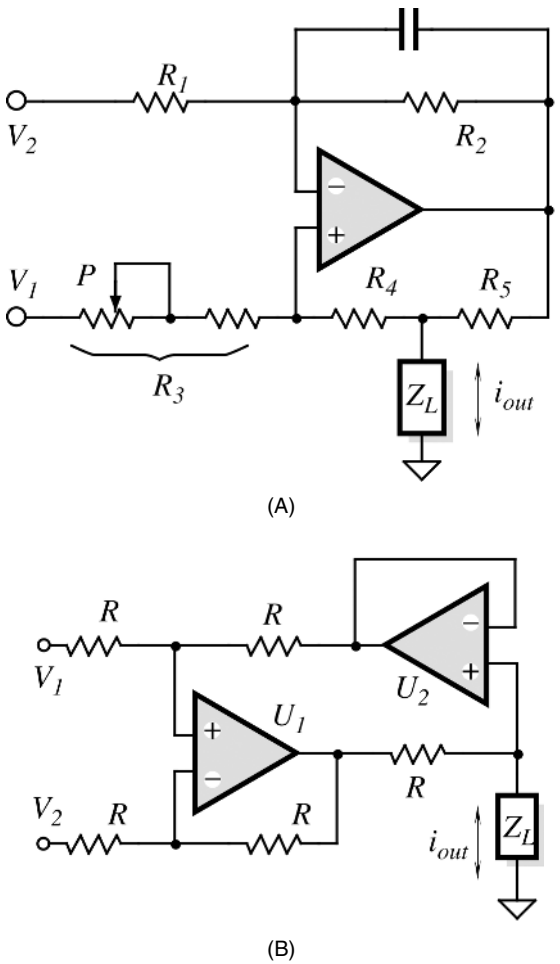


Fig. 5.15. Current generators with ground-referenced loads: (A) Howland current pump; (B) current pump with two OPAMs.

An efficient current pump with four matched resistors and two operational amplifiers is shown in Fig. 5.15B. Its output current is defined by the equation

$$i_{\text{out}} = \frac{(V_1 - V_2)}{R_s}. \tag{5.20}$$

The advantage of this circuit is that resistors R may be selected with a relatively high value and housed in the same thermally homogeneous packaging for better thermal tracking.

For the generation of low-level constant currents, a monolithic voltage reference shown in Fig. 5.16 may be found quite useful. The circuit contains a 2.5 V reference which is powered by the output current from the voltage follower U_1. The voltage

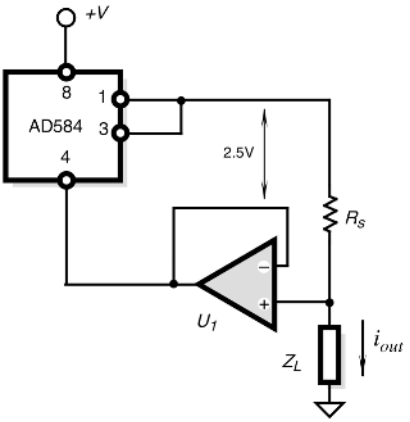


Fig. 5.16. Current source with a monolithic voltage reference.

regulator keeps the voltage drop across R_s precisely equal to 2.5 V, thus assuring that the current through that resistor and, subsequently, through the load is constant. The load current is defined as

$$i_{\text{out}} = \frac{2.5\ \text{V}}{R_s}. \tag{5.21}$$

5.3.2 Voltage References

A voltage reference is an electronic device which generates constant voltage that is little affected by variations in power supply, temperature, load, aging, and other factors. There are several techniques known for the generation of such voltages. Many voltage references are available in monolithic forms; however, in low-cost applications, especially in consumer products, a simple device known as a *zener diode* is often employed.

A zener diode has a constant voltage drop in a circuit when provided with a fairly constant current derived from a higher voltage elsewhere within the circuit. The active portion of a zener diode is a reverse-biased semiconductor p-n junction. When the diode is forward biased (the p region is more positive), there is little resistance to current flow. Actually, a forward-biased zener diode looks very much like a conventional semiconductor diode (Fig. 5.17A). When the diode is reverse biased (minus is at the anode and plus is at the cathode), very little current flows through it if the applied voltage is less than V_z. A reverse saturation current is a small leakage which is almost independent of the applied voltage. When the reverse voltage approaches the breakdown voltage V_z, the reverse current increases rapidly and, if not limited, will result in the diode overheating and destruction. For that reason, zener diodes usually are used with current-limiting components, such as resistors, positive temperature coefficient (PTC) thermistors or current sources. The most common circuit with a zener diode is shown in Fig. 5.17B. In this circuit, the diode is connected in parallel with its load, thus implying the name "shunting reference". If the load consumes current i_L, the current i should be sufficiently high to be shared between the zener diode (i_z) and the load. The zener voltage decreases as temperature of the junction rises.

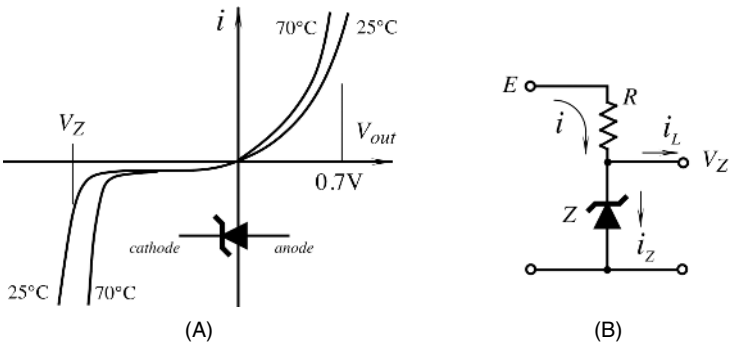


Fig. 5.17. Zener diode: (A) a volt–ampere characteristic; (B) a shunting-type voltage reference.

Zener diodes fall into three general classifications: regulator diodes, reference diodes, and transient voltage suppressors. Regulator diodes are normally employed in power supplies where a nearly constant dc output voltage is required despite relatively large changes in input voltage or load impedance. Such devices are available with a wide range of voltage and power ratings, making them suitable for a wide variety of electronic equipments. Regulator diodes, however, have one limitation: they are temperature sensitive. Therefore, in applications in which the output voltage must remain within narrow limits during input-voltage, load-current, and temperature changes, a temperature-compensated regulator diode, called a reference diode, is required.

It makes sense to take advantage of the differing thermal characteristics of forward- and reverse-biased silicon p-n junctions. Like any silicon diode (see Section 16.3 of Chapter 16), a forward-biased junction has a negative temperature coefficient of approximately -2 mV/°C, whereas a reverse-biased junction has positive temperature coefficient ranging from about 2 to 6 mV/°C depending on the current and the diode type. Therefore, it is possible, by a selective combination of forward- and reverse-biased junctions, to fabricate a device with a very low overall temperature coefficient (Fig. 5.18). The voltage changes of the two junctions are equal and opposite only at the specified current. For any other value of current, the temperature compensation may not be ideally accomplished. Nevertheless, even a simple back-to-back connection of two zener diodes of the same type may significantly improve the overall temperature stability over a rather broad range of currents and temper-

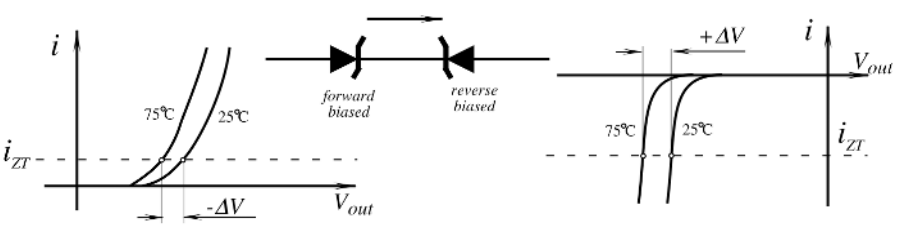


Fig. 5.18. Temperature compensation of a zener diode.

atures. Naturally, the reference voltage from the combination is higher than from a single zener diode. A commercial monolithic version of this circuit is available from Motorola (part 1N821).

The so-called band-gap references are often useful substitutes for zener diodes. They have typically 10 times lower output impedance than low-voltage zeners and can be obtained in a variety of nominal output voltages, ranging from 1.2 to 10 V. Currently, a large variety of high-quality voltage references with selectable outputs is available from many manufacturers.

5.3.3 Oscillators

Oscillators are generators of variable electrical signals. Any oscillator is essentially comprised of a circuit with a gain stage, some nonlinearity and a certain amount of positive feedback. By definition, an oscillator is an unstable circuit (as opposed to an amplifier which better be stable!) whose timing characteristics should be either steady or changeable according to a predetermined functional dependence. The latter is called a *modulation*. Generally, there are three types of electronic oscillators classified according to the time-keeping components: the RC, the LC, and the crystal oscillators. In the RC oscillators, the operating frequency is defined by capacitors (C) and resistors (R); in the LC oscillators, it is defined by the capacitive (C) and inductive (L) components. In the crystal oscillators, the operating frequency is defined by a mechanical resonant in specific cuts of piezoelectric crystals, usually quartz and ceramics.

There is a great variety of oscillation circuits, the coverage of which is beyond the scope of this book. Below, we briefly describe some practical circuits which can be used for either direct interface with sensors or may generate excitation signals in an economical fashion.

Many various multivibrators can be built with logic circuits, (e.g., with NOR, NAND gates, or binary inverters). Also, many multivibrators can be designed with comparators or operational amplifiers having a high open-loop gain. In all of these oscillators, a capacitor is being charged, and the voltage across it is compared with another voltage, which is either constant or changing. The moment when both voltages are equal is detected by a comparator. A comparator is a two-input circuit which generates an output signal when its input signals are equal. A comparator is a nonlinear element due to its very high gain that essentially results in the saturation of its output when the input signals differ by a relatively small amount. The indication of a comparison is fed back to the RC network to alter the capacitor charging in the opposite direction, which is discharging. Recharging in a new direction goes on until the next moment of comparison. This basic principle essentially requires the following minimum components: a capacitor, a charging circuit, and a threshold device (a comparator). Several relaxation oscillators are available from many manufacturers, (e.g., a very popular timer, type 555, which can operate in either monostable or actable modes). For illustration, below we describe just two discrete-component square-wave oscillators; however, there is a great variety of such circuits, which the reader can find in many books on operational amplifiers and digital systems, (e.g., Ref. [3]).

A simple square-wave oscillator can be built with two logic inverters (e.g., CMOS) (Fig. 5.19A). A logic inverter has a threshold near a half of the power supply-voltage.

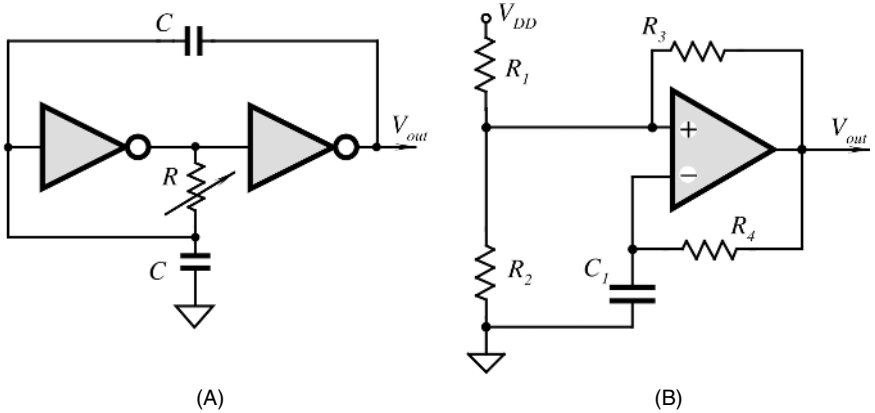


Fig. 5.19. Square-wave oscillators: (A) with two logic inverters; (B) with a comparator or OPAM.

When the voltage at its input crosses the threshold, the inverter generates the output signal of the opposite direction; that is, if the input voltage is ramping up, at the moment when it reaches one-half of the power supply, the output voltage will be a negative-going transient. Timing properties of the oscillator are determined by the resistor R and the capacitor C. Both capacitors should be of the same value. Stability of the circuit primarily depends on the stabilities of the R and C.

A very popular square-wave oscillator (Fig. 5.19B) can be built with one OPAM or a voltage comparator.[2] The amplifier is surrounded by two feedback loops: one is negative (to an inverting input) and the other is positive (to a noninverting input). The positive feedback (R_3) controls the threshold level, and the negative loop charges and discharges the timing capacitor C_1, through the resistor R_4. The frequency of this oscillator can be determined from

$$f = \frac{1}{R_4 C_1}\left[\ln\left(1 + \frac{R_1||R_2}{R_3}\right)\right]^{-1}, \tag{5.22}$$

where $R_1||R_2$ is an equivalent resistance of parallel-connected R_1 and R_2.

The two circuits (A and B) shown in Fig. 5.20 can generate sine-wave signals. They use the n-p-n transistors as amplifiers and the LC networks to set the oscillating frequency. The (B) circuit is especially useful for driving the linear variable differential transformer (LVDT) position sensors, as the sensor's transformer becomes a part of the oscillating circuit.

A radio-frequency oscillator can be used as a part of a capacitive occupancy detector to detect the presence of people in the vicinity of its antenna (Fig. 5.21).[3] The antenna is a coil which together with the C_2 capacitors determines the oscillating frequency. The antenna is coupled to the environment through its distributed capaci-

[2] A voltage comparator differs from an operational amplifier by its faster speed response and the output circuit which is easier interfaceable with TTL and CMOS logic.

[3] See Section 7.3 of Chapter 7.

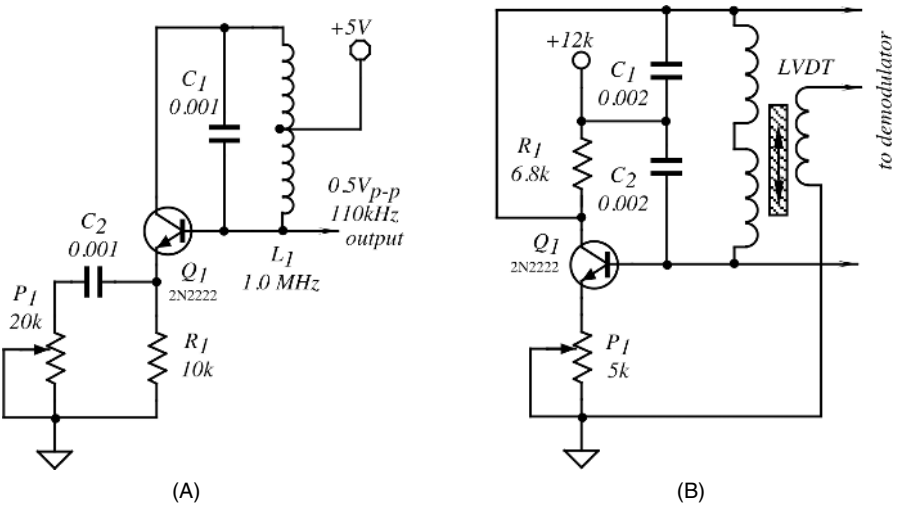


Fig. 5.20. LC sine-wave oscillators.

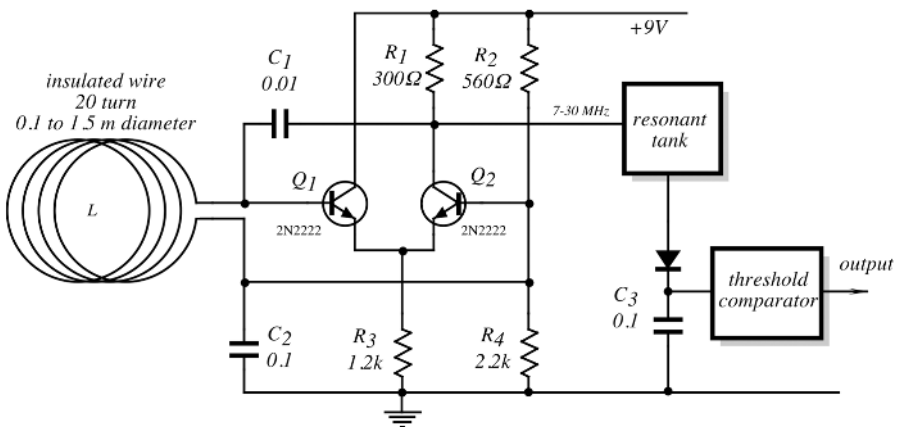


Fig. 5.21. LC radio-frequency oscillator as a capacitive occupancy detector.

tance, which somewhat reduces the frequency of the oscillator. When a person moves into the vicinity of the antenna, he/she brings in an additional capacitance which lowers the oscillator frequency even further. The output of the oscillator is coupled to a resonant tank (typically, an LC network) tuned to a baseline frequency (near 30 MHz). A human intrusion lowers the frequency, thus substantially reducing the amplitude of the output voltage from the tank. The high-frequency signal is rectified by a peak detector and a low-frequency voltage is compared with a predetermined threshold by a comparator. This circuit employs an oscillator whose frequency is modulated by a sensing antenna. However, for other applications, the same circuit with a small inductor instead of an antenna can produce stable sinusoidal oscillations.

5.3.4 Drivers

As opposed to current generators, voltage drivers must produce output voltages which, over broad ranges of the loads and operating frequencies, are independent of the output currents. Sometimes, the drivers are called hard-voltage sources. Usually, when the sensor which has to be driven is purely resistive, a driver can be a simple output stage which can deliver sufficient current. However, when the load contains capacitances or inductances (i.e., the load is reactive), the output stage becomes a more complex device.

In many instances, when the load is purely resistive, there still can be some capacitance associated with it. This may happen when the load is connected though lengthy wires or coaxial cables. A coaxial cable behaves as a capacitor connected from its central conductor to its shield if the length of the cable is less than one-fourth of the wavelength in the cable at the frequency of interest f. For a coaxial cable, this maximum length is given by

$$L \leq 0.0165 \frac{c}{f}, \tag{5.23}$$

where c is the velocity of light in a coaxial cable dielectric.

For instance, if $f = 100$ kHz, $L \leq 0.0165(3 \times 10^8/10^5) = 49.5$; that is, a cable less than 49.5 m (162.4 ft) long will behave as a capacitor connected in parallel with the load (Fig. 5.22A). For an R6-58A/U cable, the capacitance is 95 pF/m. This capacitance must be considered for two reasons: the speed and stability of the circuits. The instability results from the phase shift produced by the output resistance of the driver R_0 and the loading capacitance C_L:

$$\varphi = \arctan(2\pi f R_0 C_L). \tag{5.24}$$

For instance, for $R_0 = 100\Omega$ and $C_L = 1000$ pF, at $f = 1$ MHz, the phase shift $\varphi \approx 32°$. This shift significantly reduces the phase margin in a feedback network which may cause a substantial degradation of the response and a reduced ability to drive capacitive loads. The instability may be either overall, when an entire system oscillates, or localized when the driver alone becomes unstable. The local instabilities often can be

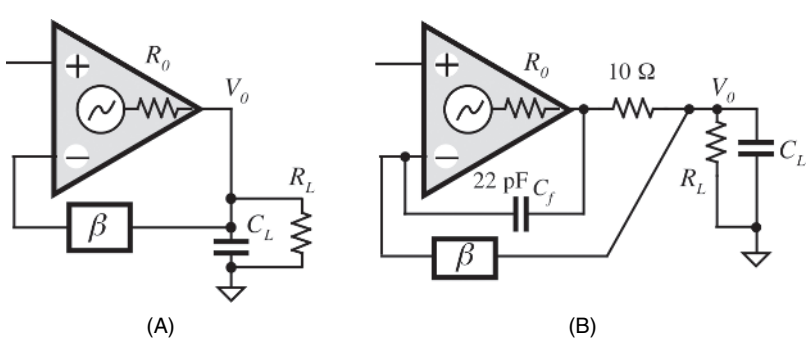


Fig. 5.22. Driving a capacitive load: (A) a load capacitor is coupled to the driver's input through a feedback; (B) decoupling of a capacitive load.

cured by large bypass capacitors (on the order of 10 μF) across the power supply or the so-called Q-spoilers consisting of a serial connection of a 3–10Ω resistor and a disk ceramic capacitor connected from the power-supply pins of the driver chip to ground.

To make a driver stage more tolerant to capacitive loads, it can be isolated by a small serial resistor, as is shown in Fig. 5.22B. A small capacitive feedback (C_f) to the inverting input of the amplifier and a 10Ω resistor may allow driving loads as large as 0.5 μF. However, in any particular case, it is recommended to find the best values for the resistor and the capacitor experimentally.

5.4 Analog-to-Digital Converters

5.4.1 Basic Concepts

The analog-to-digital (A/D) converters range from discrete circuits, to monolithic ICs (integrated circuits), to high-performance hybrid circuits, modules, and even boxes. Also, the converters are available as standard cells for custom and semicustom application-specific integrated circuits (ASICs). The A/D converters transform analog data—usually voltage—into an equivalent digital form, compatible with digital data processing devices. Key characteristics of A/D converters include absolute and relative accuracy, linearity, no missing codes, resolution, conversion speed, stability, and price. Quite often, when price is of a major concern, discrete-component or monolithic IC versions are the most efficient. The most popular A/D converters are based on a successive-approximation technique because of an inherently good compromise between speed and accuracy. However, other popular techniques are used in a large variety of applications, especially when a high conversion speed is not required. These include dual-ramp, quad-slope, and voltage-to-frequency (V/F) converters. The art of an A/D conversion is well developed. Here, we briefly review some popular architectures of the converters; however, for detailed descriptions the reader should refer to specialized texts, such as Ref. [4].

The best known digital code is *binary* (base 2). Binary codes are most familiar in representing integers; that is, in a natural binary integer code having n bits, the LSB (least significant bit) has a weight of 2 (i.e., 1), the next bit has a weight of 2^1 (i.e., 2), and so on up to MSB (most significant bit), which has a weight of 2^{n-1} (i.e., $2^n/2$). The value of a binary number is obtained by adding up the weights of all nonzero bits. When the weighted bits are added up, they form a unique number having any value from 0 to $2^n - 1$. Each additional trailing zero bit, if present, essentially doubles the size of the number.

When converting signals from analog sensors, because the full scale is independent of the number of bits of resolution, a more useful coding is *fractional* binary [4], which is always normalized to full scale. Integer binary can be interpreted as fractional binary if all integer values are divided by 2^n. For example, the MSB has a weight of 1/2 (i.e., $2^{n-1}/2^n = 2^{-1}$), the next bit has a weight of 1/4 (i.e., 2^{-2}), and so forth down to the LSB, which has a weight of $1/2^n$ (i.e., 2^{-n}). When the weighted bits are added up, they form a number with any of 2^n values, from 0 to $(1 - 2^{-n})$ of full scale.

Table 5.1. Integer and Fractional Binary Codes

Decimal Fraction	Binary Fraction	MSB x1/2	Bit2 x1/4	Bit3 x1/6	Bit4 x1/16	Binary Integer	Decimal Integer
0	0.0000	0	0	0	0	0000	0
1/16 (LSB)	0.0001	0	0	0	1	0001	1
$2/16 = 1/8$	0.0010	0	0	1	0	0010	2
$3/16 = 1/8 + 1/16$	0.0011	0	0	1	1	0011	3
$4/16 = 1/4$	0.0100	0	1	0	0	0100	4
$5/16 = 1/4 + 1/16$	0.0101	0	1	0	1	0101	5
$6/16 = 1/4 + 1/8$	0.0110	0	1	1	0	0110	6
$7/16 = 1/4 + 1/8 + 1/16$	0.0111	0	1	1	1	0111	7
$8/16 = 1/2$ (MSB)	0.1000	1	0	0	0	1000	8
$9/16 = 1/2 + 1/16$	0.1001	1	0	0	1	1001	9
$10/16 = 1/2 + 1/8$	0.1010	1	0	1	0	1010	10
$11/16 = 1/2 + 1/8 + 1/16$	0.1011	1	0	1	1	1011	11
$12/16 = 1/2 + 1/4$	0.1100	1	1	0	0	1100	12
$13/16 = 1/2 + 1/4 + 1/16$	0.1101	1	1	0	1	1101	13
$14/16 = 1/2 + 1/4 + 1/8$	0.1110	1	1	1	0	1110	14
$15/16 = 1/2 + 1/4 + 1/8 + 1/16$	0.1111	1	1	1	1	1111	15

Source: Adapted from Ref. [4].

Additional bits simply provide more fine structure without affecting the full-scale range. To illustrate these relationships, Table 5.1 lists 16 permutations of 5-bit's worth of 1's and 0's, with their binary weights, and the equivalent numbers expressed as both decimal and binary integers and fractions.

When all bits are "1" in natural binary, the fractional number value is $1 - 2^{-n}$, or normalized full-scale less 1 LSB ($1 - 1/16 = 15/16$ in the example). Strictly speaking, the number that is represented, written with an "integer point," is 0.1111 ($= 1 - 0.0001$). However, it is almost universal practice to write the code simply as the integer 1111 (i.e., "15") with the fractional nature of the corresponding number understood: "1111" $\rightarrow$ $1111/(1111 + 1)$, or 15/16.

For convenience, Table 5.2 lists bit weights in binary for numbers having up to 20 bits. However, the practical range for the vast majority of sensors rarely exceeds 16 bits.

The weight assigned to the LSB is the resolution of numbers having n bits. The dB column represents the logarithm (base 10) of the ratio of the LSB value to unity [full scale (FS)], multiplied by 20. Each successive power of 2 represents a change of 6.02 dB [i.e., $20 \log_{10}(2)$] or "6 dB/octave."

5.4.2 V/F Converters

The voltage-to-frequency (V/F) converters can provide a high-resolution conversion, and this is useful for the sensor's special features as a long-term integration (from seconds to years), a digital-to-frequency conversion (together with a D/A converter), a frequency modulation, a voltage isolation, and an arbitrary frequency division and

Table 5.2. Binary Bit Weights and Resolutions

Bit	2^{-n}	$1/2^n$ Fraction	dB	$1/2^n$ Decimal	%	ppm
FS	2^0	1	0	1.0	100	1,000,000
MSB	2^{-1}	1/2	−6	0.5	50	500,000
2	2^{-2}	1/4	−12	0.25	25	250,000
3	2^{-3}	1/8	−18.1	0.125	12.5	125,000
4	2^{-4}	1/16	−24.1	0.0625	6.2	62,500
5	2^{-5}	1/32	−30.1	0.03125	3.1	31,250
6	2^{-6}	1/64	−36.1	0.015625	1.6	15,625
7	2^{-7}	1/128	−42.1	0.007812	0.8	7,812
8	2^{-8}	1/256	−48.2	0.003906	0.4	3,906
9	2^{-9}	1/512	−54.2	0.001953	0.2	1,953
10	2^{-10}	1/1,024	−60.2	0.0009766	0.1	977
11	2^{-11}	1/2,048	−66.2	0.00048828	0.05	488
12	2^{-12}	1/4,096	−72.2	0.00024414	0.024	244
13	2^{-13}	1/8,192	−78.3	0.00012207	0.012	122
14	2^{-14}	1/16,384	−84.3	0.000061035	0.006	61
15	2^{-15}	1/32,768	−90.3	0.0000305176	0.003	31
16	2^{-16}	1/65,536	−96.3	0.0000152588	0.0015	15
17	2^{-17}	1/131,072	−102.3	0.00000762939	0.0008	7.6
18	2^{-18}	1/262,144	−108.4	0.000003814697	0.0004	3.8
19	2^{-19}	1/524,288	−114.4	0.000001907349	0.0002	1.9
20	2^{-20}	1/1,048,576	−120.4	0.0000009536743	0.0001	0.95

multiplication. The converter accepts an analog output from the sensor, which can be either voltage or current (in the latter case, of course, it should be called a current-to-frequency converter). In some cases, a sensor may become a part of an A/D converter, as is illustrated in Section 5.5. Here, however, we will discuss only the conversion of voltage to frequency, or, in other words, to a number of square pulses per unit of time. The frequency is a digital format because pulses can be gated (selected for a given interval of time) and then counted, resulting in a binary number. All V/F converters are of the integrating type because the number of pulses per second, or *frequency*, is proportional to the *average* value of the input voltage.

By using a V/F converter, an A/D can be performed in the most simple and economical manner. The time required to convert an analog voltage into a digital number is related to the full-scale frequency of the V/F converter and the required resolution. Generally, the V/F converters are relatively slow, as compared with successive-approximation devices; however, they are quite appropriate for the vast majority of sensor applications. When acting as an A/D converter, the V/F converter is coupled to a counter which is clocked with the required sampling rate. For instance, if a full-scale frequency of the converter is 32 kHz and the counter is clocked eight times per second, the highest number of pulses which can be accumulated every counting cycle is 4000 which approximately corresponds to a resolution of 12 bits (see Table 5.2). By using the same combination of components (the V/F converter and the counter), an

integrator can be built for the applications, where the stimulus needs to be integrated over a certain time. The counter accumulates pulses over the gated interval rather than as an average number of pulses per counting cycle.

Another useful feature of a V/F converter is that its pulses can be easily transmitted through communication lines. The pulsed signal is much less susceptible to a noisy environment than a high-resolution analog signal. In the ideal case, the output frequency f_{out} of the converter is proportional to the input voltage V_{in}:

$$\frac{f_{\text{out}}}{f_{\text{FS}}} = \frac{V_{\text{in}}}{V_{\text{FS}}}, \tag{5.25}$$

where f_{FS} and V_{FS} are the full-scale frequency and input voltage, respectively. For a given linear converter, ratio $f_{\text{FS}}/V_{\text{FS}} = G$ is constant and is called a conversion factor; then,

$$f_{\text{out}} = G V_{\text{in}}. \tag{5.26}$$

There are several known types of V/F converters. The most popular of them are the multivibrator and the charge-balance circuit.

A *multivibrator V/F converter* employs a free-running square-wave oscillator where charge–discharge currents of a timing capacitor are controlled by the input signal (Fig. 5.23). The input voltage V_{in} is amplified by a differential amplifier (e.g., an instrumentation amplifier) whose output signal controls two voltage-to-current converters (transistors U_1 and U_2). A precision multivibrator alternatively connects timing capacitor C to both current converters. The capacitor is charged for a half of period through transistor U_1 by the current i_a. During the second half of the timing period, it is discharged by the current i_b through transistor U_2. Because currents i_a and i_b are controlled by the input signal, the capacitor charging and discharging slopes vary accordingly, thus changing the oscillating frequency. An apparent advantage of this circuit is its simplicity and potentially very low power consumption; however, its ability to reject high-frequency noise in the input signal is not as good as in the charge-balance architecture.

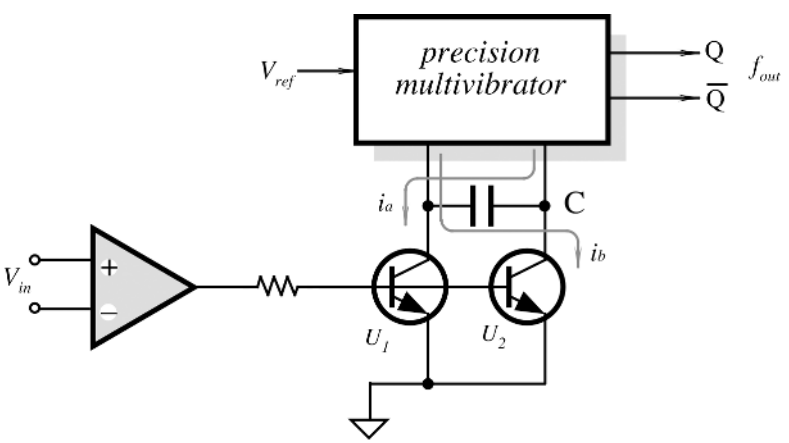


Fig. 5.23. Multivibrator type of voltage-to-frequency converter.

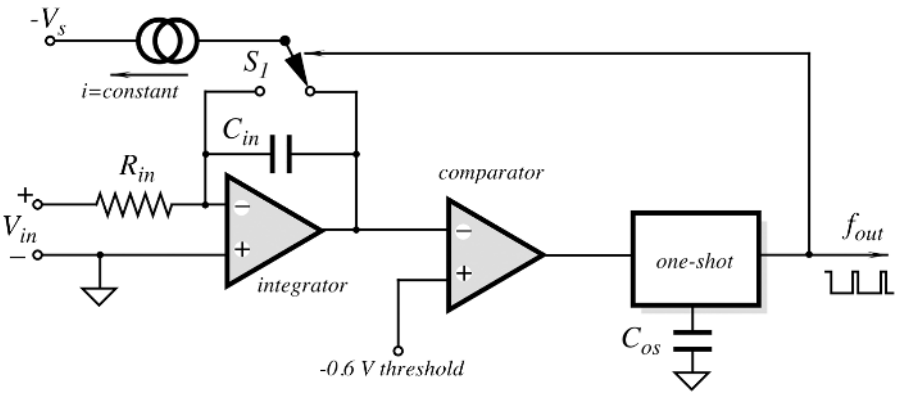


Fig. 5.24. Charge-balance V/F converter.

The *charge-balance* type of converter employs an analog integrator and a voltage comparator, as shown in Fig. 5.24. This circuit has such advantages as high speed, high linearity, and good noise rejection. The circuit is available in an integral form from several manufacturers—for instance, ADVFC32 and AD650 from Analog Devices, and LM331 from National Semiconductors. The converter operates as follows. The input voltage V_{in} is applied to an integrator through the input resistor R_{in}. The integrating capacitor is connected as a negative feedback loop to the operational amplifier whose output voltage is compared with a small negative threshold of -0.6 V. The integrator generates a saw-tooth voltage (Fig. 5.26), which, at the moment of comparison with the threshold, results in a transient at the comparator's output. That transient enables a one-shot generator which produces a square pulse of a fixed duration t_{os}. A precision current source generates constant current i, which is alternatively applied either to the summing node of the integrator or to its output. The switch S_1 is controlled by the one-shot pulses. When the current source is connected to the summing node, it delivers a precisely defined packet of charge $\Delta Q = i t_{\text{os}}$ to the integrating capacitor. The same summing node also receives an input charge through the resistor R_{in}, thus the net charge is accumulated on the integrating capacitor C_{in}.

When the threshold is reached, the one-shot is triggered and the switch S_1 changes its state to high, thus initiating a reset period (Fig. 5.25B). During the reset period, the current source delivers its current to the summing node of the integrator. The input current charges the integrating capacitor upward. The total voltage between the threshold value and the end of the deintegration is determined by the duration of a one-shot pulse:

$$\Delta V = t_{\text{os}} \frac{dV}{dt} = t_{\text{os}} \frac{i - I_{\text{in}}}{C_{\text{in}}}. \tag{5.27}$$

When the output signal of the one-shot circuit goes low, switch S_1 diverts current i to the output terminal of an integrator, which has no effect on the state of the integrating capacitor C_{in}; that is, the current source sinks a portion of the output current from the operational amplifier. This time is called the integration period (Figs. 5.25A and 5.26). During the integration, the positive input voltage delivers current $I_{\text{in}} = V_{\text{in}}/R_{\text{in}}$ to the

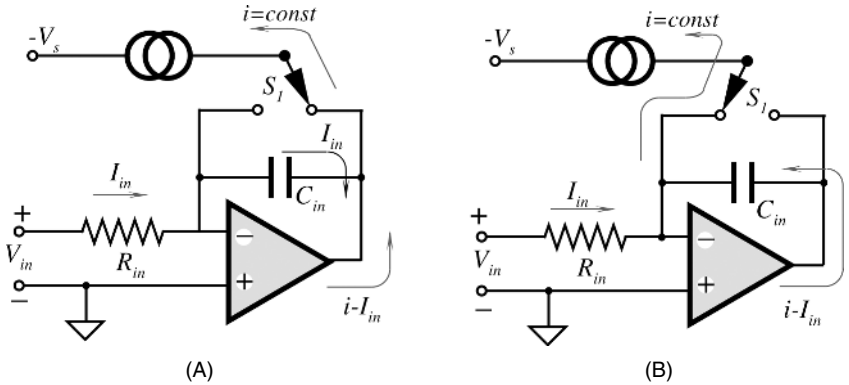


Fig. 5.25. Integrate and deintegrate (reset) phases in a charge-balance converter.

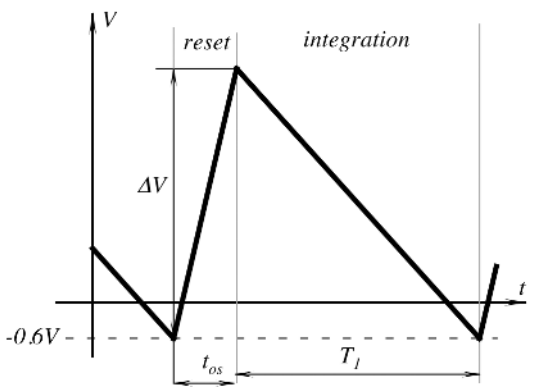


Fig. 5.26. Integrator output in a charge-balance converter.

capacitor C_{in}. This causes the integrator to ramp down from its positive voltage with the rate proportional to V_{in}. The amount of time required to reach the comparator's threshold is

$$T_1 = \frac{\Delta V}{dV/dt} = t_{\text{os}} \frac{i - I_{\text{in}}}{C_{\text{in}}} \frac{1}{I_{\text{in}}/C_{\text{in}}} = t_{\text{os}} \frac{i - I_{\text{in}}}{I_{\text{in}}}. \tag{5.28}$$

It is seen that the capacitor value does not affect the duration of the integration period. The output frequency is determined by

$$f_{\text{out}} = \frac{1}{t_{\text{os}} + T_1} = \frac{I_{\text{in}}}{t_{\text{os}} i} = \frac{V_{\text{in}}}{R_{\text{in}}} \frac{1}{t_{\text{os}} i}. \tag{5.29}$$

Therefore, the frequency of one-shot pulses is proportional to the input voltage. It depends also on the quality of the integrating resistor, stability of the current generator, and a one-shot circuit. With a careful design, this type of V/F converter may reach a nonlinearity error below 100 ppm and can generate frequencies from 1 Hz to 1 MHz.

A major advantage of the integrating-type converters, such as a charge-balanced V/F converter, is the ability to reject large amounts of additive noise; by integrating the measurement, noise is reduced or even totally eliminated. Pulses from the converter are accumulated for a gated period T in a counter. Then, the counter behaves like a filter having a transfer function in the form

$$H(f) = \frac{\sin \pi f T}{\pi f T}, \tag{5.30}$$

where f is the frequency of pulses. For low frequencies, the value of this transfer function $H(f)$ is close to unity, meaning that the converter and the counter make correct measurements. However, for a frequency $1/T$, the transfer function $H(1/T)$ is zero, meaning that these frequencies are completely rejected. For example, if the gating time $T = 16.67$ ms which corresponds to a frequency of 60 Hz (the power line frequency which is a source of substantial noise in many sensors), then the 60 Hz noise will be rejected. Moreover, the multiple frequencies (120 Hz, 180 Hz, 240 Hz, and so on) will also be rejected.

5.4.3 Dual-Slope Converter

A dual-slope converter is very popular; it is used nearly universally in handheld digital voltmeters and other portable instruments where a fast conversion is not required. This type of converter performs an indirect conversion of the input voltage. First, it converts V_{in} into a function of time; then, the time function is converted into a digital number by a pulse counter. Dual-slope converters are quite slow; however, for stimuli which do not exhibit fast changes, they are often the converters of choice, due to their simplicity, cost-effectiveness, noise immunity, and potentially high resolution. The operating principle of the converter is as follows (Fig. 5.27). Like in a charge-balance converter, there is an integrator and a threshold comparator. The threshold level is set at zero (ground) or any other suitable constant voltage. The integrator can be selectively connected through the analog selector S_1 either to the input voltage or to the reference voltage. In this simplified schematic, the input voltage is negative, and the reference voltage is positive. However, by shifting the dc level of the input signal (with the help of an additional OPAM), the circuit will be able to convert bipolar input signals as well. The output of the comparator sends a signal to the control logic when the integrator's output voltage crosses zero. The logic controls both the selector S_1 and the reset switch S_2, which serves for discharging the integrating capacitor, C_{in}.

When the start input is enabled, S_1 connects the integrator to the input signal and the logic starts a timer. The timer is preset for a fixed time interval T. During that time, the integrator generates a positive-going ramp (Fig. 5.28), which changes according to the input signal. It should be noted that the input signal does not have to be constant. Any variations in the signal are averaged during the integration process. Upon elapsing time T, the integrator output voltage reaches the level

$$V_m = \overline{V}_{\text{in}} \frac{T}{R_{\text{in}} C_{\text{in}}}, \tag{5.31}$$

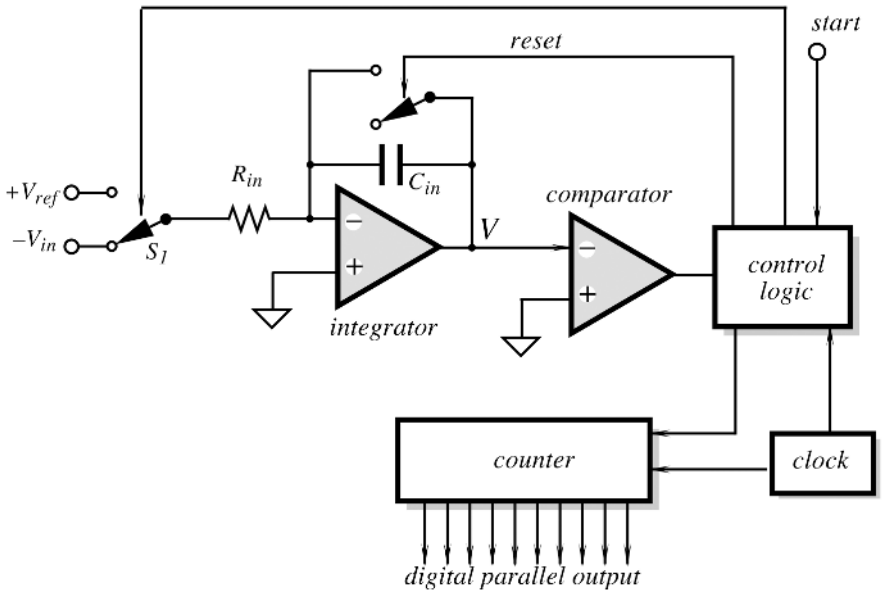


Fig. 5.27. Dual-slope A/D converter.

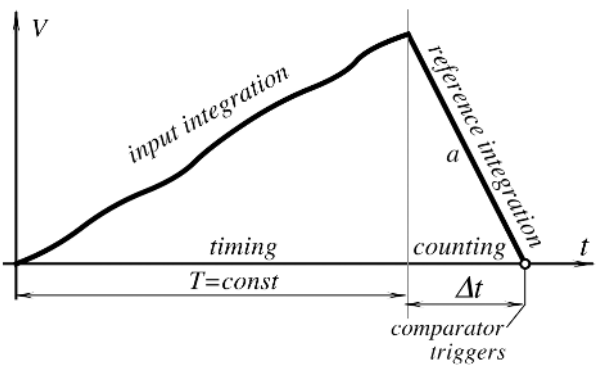


Fig. 5.28. Integrator output in a dual-slope A/D converter.

where $\overline{V}_{\text{in}}$ is an average input signal during time T. At that moment, S_1 switches to the reference voltage which is of the opposite polarity with respect to the input signal, thus setting the deintegrate phase (a reference voltage integration), during which the integrator's voltage ramps downward until it crosses a zero threshold. The integral of the reference has slope

$$a = -\frac{V_{\text{ref}}}{R_{\text{in}} C_{\text{in}}}. \tag{5.32}$$

During the deintegrate phase, the counter counts clock pulses. When the comparator indicates a zero crossing, the count is stopped and the analog integrator is reset by

discharging its capacitor through S_2. The charge at the capacitor gained during the input signal integrate phase is precisely equal to the charge lost during the reference deintegration phase. Therefore, the following holds:

$$\overline{V}_{\text{in}} \frac{T}{R_{\text{in}} C_{\text{in}}} = V_{\text{ref}} \frac{\Delta t}{R_{\text{in}} C_{\text{in}}}, \tag{5.33}$$

which leads to

$$\frac{\overline{V}_{\text{in}}}{V_{\text{ref}}} = \frac{\Delta t}{T}. \tag{5.34}$$

Therefore, the ratio of the average input voltage and the reference voltage is replaced by the ratio of two time intervals. Then, the counter does the next step—it converts the time interval Δt into a digital form by counting the clock pulses during Δt. The total count is the measure of $\overline{V}_{\text{in}}$ (remember that V_{ref} and T are constants).

The dual-slope converter has the same advantage as the charge-balance converter: They both reject frequencies $1/T$ corresponding to the integrate timing. It should be noted that selecting time $T = 200$ ms will reject noise produced by both 50 and 60 Hz, thus making the converter immune to the power line noise originated at either standard frequency. Further, the conversion accuracy is independent of the clock frequency stability, because the same clock sets timing T and the counter. The resolution of the conversion is limited only by the analog resolution; hence, the excellent fine structure of the signal may be represented by more bits than would be needed to maintain a given level of scale-factor accuracy. The integration provides rejection of high-frequency noise and averages all signal instabilities during the interval T. The throughput of a dual-slope conversion is limited to somewhat less than $1/2T$ conversions per second.

To minimize errors produced in the analog portion of the circuit (the integrator and the comparator), a third timing phase is usually introduced. It is called an auto-zero phase because during that phase, the capacitor C_{in} is charged with zero-drift errors, which are then introduced in the opposite sense during the integration, in order to nullify them. An alternative way to reduce the static error is to store the auto-zero counts and then digitally subtract them.

Dual-slope converters are often implemented as a combination of analog components (OPAMs, switches, resistors, and capacitors) and a microcontroller, which handles the functions of timing, control logic, and counting. Sometimes, the analog portion is packaged in a separate integrated circuit. An example is the TS500 module from Texas Instruments.

5.4.4 Successive-Approximation Converter

These converters are widely used in a monolithic form thanks to their high speed (to 1 MHz throughput rates) and high resolution (to 16 bits). The conversion time is fixed and independent of the input signal. Each conversion is unique, as the internal logic and registers are cleared after each conversion, thus making these A/D converters suitable for the multichannel multiplexing. The converter (Fig. 5.29A) consists of a precision voltage comparator, a module comprising shifter registers and a control

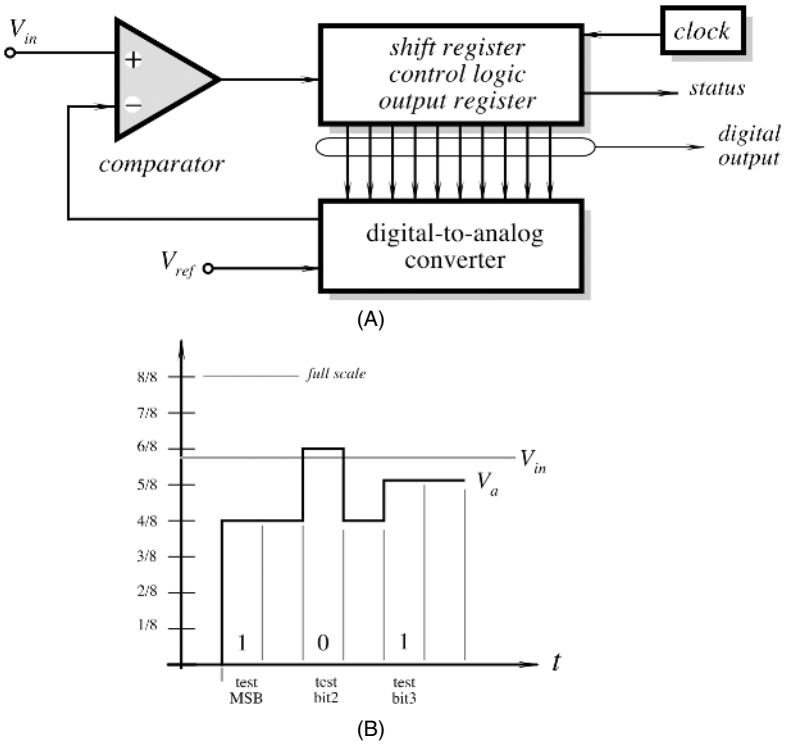


Fig. 5.29. Successive-approximation A/D converter: (A) block diagram; (B) 3-bit weighing.

logic, and a digital-to-analog converter (D/A) which serves as a feedback from the digital outputs to the input analog comparator.

The conversion technique consists of comparing the unknown input, V_{in}, against a precise voltage, V_a, or current generated by a D/A converter. The conversion technique is similar to a weighing process using a balance, with a set of n binary weights (e.g., 1/2 kg, 1/4 kg, 1/8 kg, 1/16 kg, etc. up to total of 1 kg). Before the conversion cycles, all of the registers must be cleared and the comparator's output is HIGH. The D/A converter has a MSB (1/2 scale) at its inputs and generates an appropriate analog voltage, V_a, equal to one-half of a full-scale input signal. If the input is still greater than the D/A voltage (Fig. 5.29B), the comparator remains HIGH, causing "1" at the register's output. Then, the next bit ($2/8 = 1/4$ of FS) is tried. If the second bit does not add enough weight to exceed the input, the comparator remains HIGH ("1" at the output), and the third bit is tried. However, if the second bit tips the scale too far, the comparator goes LOW, resulting in "0" in the register, and the third bit is tried. The process continues in order of descending bit weight until the last bit has been tried. After the completion, the status line indicates the end of conversion and data can be read from the register as a valid number corresponding to the input signal.

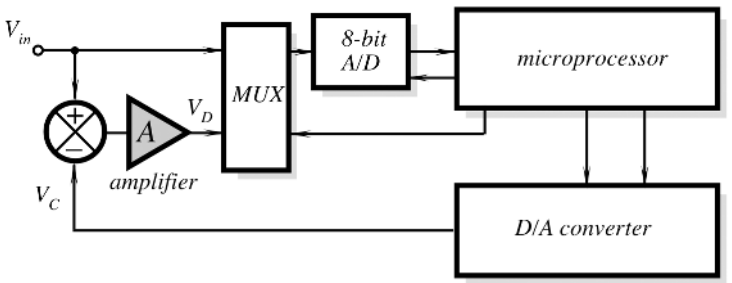


Fig. 5.30. Resolution enhancement circuit.

To make the conversion valid, the input signal V_{in} must not change until all of the bits are tried; otherwise, the digital reading may be erroneous. To avoid any problems with the changing input, a successive-approximation converter is usually supplied with a sample-and-hold (S&H) circuit. This circuit is a short-time analog memory which samples the input signal and stores it as a dc voltage during an entire conversion cycle.

5.4.5 Resolution Extension

In a typical data acquisition system, a monolithic microcontroller often contains an analog-to-digital converter, whose maximum resolution is often limited to 8 bits and sometimes to 10 bits. When the resolution is higher, 12 or even 14 bits, either the cost becomes prohibitively high or the on-the-chip A/D converter may possess several undesirable characteristics. In most applications, 8 or 10 bits may not be nearly enough for the correct representation of a stimulus. One method of achieving higher resolution is to use a dual-slope A/D converter whose resolution is limited only by the available counter rate and the speed response of a comparator.[4] Another method is to use an eight-bit A/D converter (e.g., of a successive-approximation type) with a resolution extension circuit. Such a circuit can boost the resolution by several bits, (e.g., from 8 to 12). A block diagram of the circuit is shown in Fig.5.30. In addition to a conventional eight-bit A/D converter, it includes a D/A converter, a subtraction circuit, and an amplifier having gain A. In the ASIC or discrete circuits, a D/A converter may be shared with an A/D part (see Fig. 5.29A).

The input signal V_{in} has a full-scale value E; thus for an eight-bit converter, the initial resolution will be

$$R_0 = \frac{E}{2^8 - 1} = \frac{E}{255}, \tag{5.35}$$

which is expressed in volts per bit. For instance, for a 5-V full scale, the eight-bit resolution is 19.6 mV/bit. Initially, the multiplexer (MUX) connects the input signal to the A/D converter, which produces the output digital value, M (expressed in bits). Then, the microprocessor outputs that value to a D/A converter, which produces

[4] A resolution should not be confused with accuracy.

output analog voltage V_c, which is an approximation of the input signal. This voltage is subtracted from the input signal and amplified by the amplifier to the value

$$V_D = (V_m - V_c)A. \tag{5.36}$$

The voltage V_D is an amplified error between the actual and digitally represented input signals. For a full-scale input signal, the maximum error $(V_m - V_c)$ is equal to a resolution of an A/D converter; therefore, for an eight-bit conversion $V_D = 19.6A$ mV. The multiplexer connects that voltage to the A/D converter, which converts V_D to a digital value C:

$$C = \frac{V_D}{R_0} = (V_m - V_c)\frac{A}{R_0}. \tag{5.37}$$

As a result, the microprocessor combines two digital values: M and C, where C represents the high-resolution bits. If $A = 255$, then for the 5-V full-scale, LSB $\approx$ 77 μV, which corresponds to a total resolution of 16 bits. In practice, it is hard to achieve such a high resolution because of the errors originating in the D/A converter, reference voltage, amplifier's drift, noise, and so forth. Nevertheless, the method is quite efficient when a modest resolution of 10 or 12 bits is deemed to be sufficient.

5.5 Direct Digitization and Processing

Most sensors produce low-level signals. To bring these signals to levels compatible with data processing devices, amplifiers are generally required. Unfortunately, amplifiers and connecting cables and wires may introduce additional errors, add cost to the instrument, and increase complexity. Some emerging trends in the sensor-based systems are causing use of the signal conditioning amplifiers to be reevaluated (at least for some transducers) [5]. In particular, many industrial sensor-fed systems are employing digital transmission and processing equipment. These trends point toward direct digitization of sensor outputs—a difficult task. It is especially true when a sensor-circuit integration on a single chip is considered.

Classical A/D conversion techniques emphasize high-level input ranges. This allows the LSB step size to be as large as possible, minimizing offset and noise error. For this reason, a minimum LSB signal is always selected to be at least 100–200 μV. Therefore, a direct connection of many sensors (e.g., RTD temperature transducers or piezoresistive strain gauges) is unrealistic. Such transducers' full-scale output may be limited by several millivolts, meaning that a 10-bit A/D converter must have about 1 μV LSB.

Direct digitization of transducers eliminates a dc gain stage and may yield a better performance without sacrificing accuracy. The main idea behind direct digitization is to incorporate a sensor into a signal converter, (e.g., an A/D converter or an impedance-to-frequency converter). All such converters perform a modulation process and, therefore, are nonlinear devices. Hence, they have some kind of nonlinear circuit, often a threshold comparator. Shifting the threshold level, for instance, may modulate the output signal, which is a desirable effect.

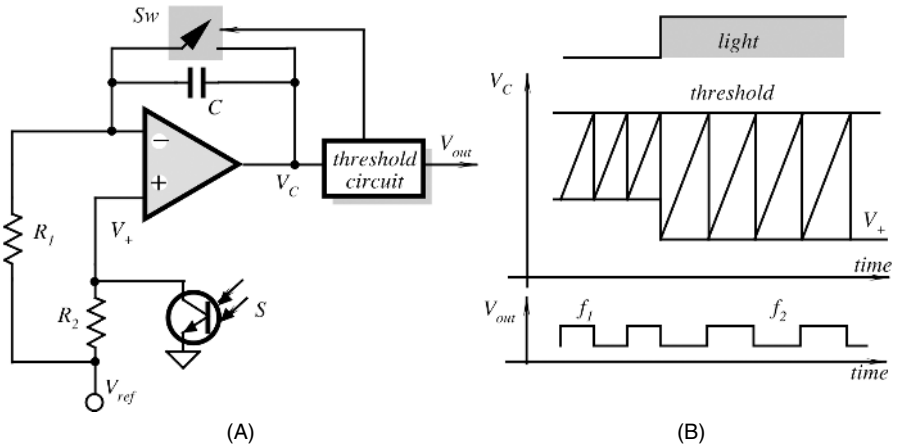


Fig. 5.31. Simplified schematic (A) and voltages (B) of a light-modulated oscillator.

Figure 5.31A shows a simplified circuit diagram of a modulating oscillator. It is composed of an integrator built with an operational amplifier and a threshold circuit. The voltage across the capacitor, C, is an integral of the current whose value is proportional to the voltage in the noninverting input of the operational amplifier. When that voltage reaches the threshold, switch SW closes, thus fully discharging the capacitor. The capacitor starts integrating the current again until the cycle repeats. The operating point of the amplifier is defined by the resistor R_2, a phototransistor S, and the reference voltage V_{ref}. A change in light flux which is incident on the base of the transistor changes its collector current, thus shifting the operation point. A similar circuit may be used for direct conversion of a resistive transducer, (e.g., a thermistor). The circuit can be further modified for accuracy enhancement, such as for the compensation of the amplifier's offset voltage or bias current, temperature drift, and so forth.

Capacitive sensors are very popular in many applications. Currently, micromachining technology allows us to fabricate small monolithic capacitive sensors. Capacitive pressure transducers employ a thin silicon diaphragm as a movable plate of the variable-gap capacitor, which is completed by a metal electrode on the opposing plate. The principal problem in these capacitors is a relatively low capacitance value per unit area (about 2 pF/mm^2) and resulting large die sizes. A typical device offers a zero-pressure capacitance on the order of few picofarads, so that an eight-bit resolution requires the detection of capacitive shifts on the order of 50 fF or less (1femtofarad $= 10^{-15}$ F). It is obvious that any external measurement circuit will be totally impractical, as parasitic capacitance of connecting conductors at best can be on the order of 1 pF—too high with respect to the capacitance of the sensor. Therefore, the only way to make such a sensor practical is to build an interface circuit as an integral part of the sensor itself. One quite effective way of designing such a circuit is to use a switched-capacitor technique. The technique is based on charge transfer from one capacitor to another by means of solid-state analog switches.

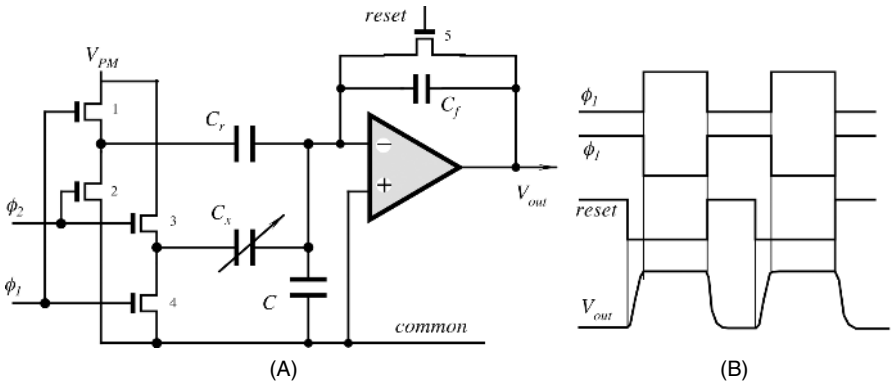


Fig. 5.32. Simplified schematic (A) and timing diagrams (B) of a differential capacitance-to-voltage converter.

Figure 5.32A shows a simplified circuit diagram of a switched-capacitor converter [6], where the variable capacitance C_x and reference capacitance C_r are parts of a symmetrical silicon pressure sensor. Monolithic MOS switches (1–4) are driven by opposite-phase clock pulses, ϕ_1 and ϕ_2. When the clocks switch, a charge appears at the common capacitance node. The charge is provided by the constant-voltage source, V_{PM}, and is proportional to $C_x - C_r$ and, therefore, to the applied pressure to the sensor. This charge is applied to a charge-to-voltage converter which includes an operational amplifier, integrating capacitor C_f, and MOS discharge (reset) switch 5. The output signal is variable-amplitude pulses (Fig. 5.32B) which can be transmitted through the communication line and either demodulated to produce a linear signal or further converted into digital data. So long as the open-loop gain of the integrating OPAM is high, the output voltage is insensitive to stray input capacitance C, offset voltage, and temperature drift. The minimum detectable signal (noise floor) is determined by the component noise and temperature drifts of the components. The circuit analysis shows that the minimum noise power occurs when the integration capacitor C_f is approximately equal to the frequency-compensation capacitor of the OPAM.

When the MOS reset switch goes from the on state to the off state, the switching signal injects some charge from the gate of the reset transistor to the input summing node of the OPAM (inverting input). This charge propagated through the gate-to-channel capacitance of the MOS transistor 5. An injection charge results in an offset voltage at the output. This error can be compensated for by a charge-canceling device [7] which can improve the signal-to-noise ratio by two orders of magnitude of the uncompensated charge. The temperature drift of the circuit can be expressed as

$$\frac{dV_{\text{out}}}{dT} = V_{\text{PM}} \frac{C_x - C_r}{C_f} (T_{C_r} - T_{C_f}), \tag{5.38}$$

where T_{C_r} is the nominal temperature coefficient of C_x and C_r, and T_{C_f} is the temperature coefficient of integrating capacitor C_f. This equation suggests that the tem-

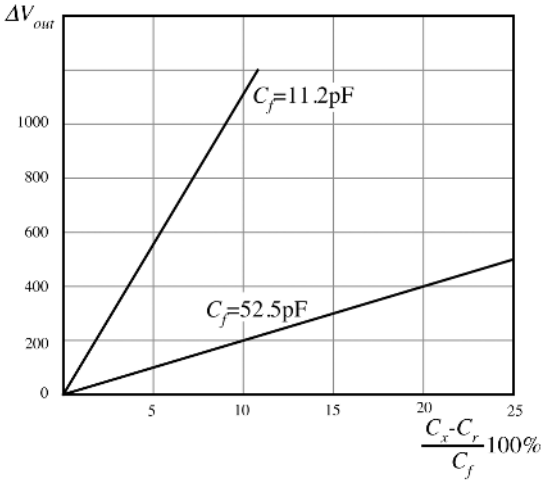


Fig. 5.33. Transfer function of a capacitance-to-voltage converter for two values of integrating capacitor. (Adapted from Ref. [6])

perature drift primarily depends on the mismatch of the capacitances in the sensor. Typical transfer functions of the circuit for two different integrating capacitors C_f are shown in Fig. 5.33. An experimental circuit similar to the above was built in a silicon die having dimensions 0.68 × 0.9 mm [8] using the standard CMOS process. The circuit operates with clock frequencies in the range from 10 to 100 kHz.

A modern trend in the sensor signal conditioning is to integrate in a single silicon chip the amplifiers, multiplexers, A/D converter, and other circuits. An example is MAX1463 (from Maxim Integrated Products); this is a highly integrated, dual-channel, 16-bit programmable sensor signal conditioner that provides amplification, calibration, signal linearization, and temperature compensation. The MAX1463 delivers an overall performance approaching the inherent repeatability of the sensor without external trim components. This circuit is designed for use with a broad range of sensors, including pressure sensing, RTD and thermocouple linearization, load cells, force sensors, and with resistive elements used in magnetic direction sensors. The MAX1463 has a choice of analog or digital outputs, including voltage, current (4–20 mA), ratiometric, and Pulse Width Modulation (PWM). Uncommitted op amps are available for buffering the digital-to-analog converter (DAC) outputs, driving heavier external loads, or providing additional gain and filtering. In addition, it incorporates a 16-bit CPU, user-programmable 4 kB of FLASH program memory, 128 bytes of FLASH user information, one 16-bit ADC, and two 16-bit DACs. It also has two 12-bit PWM digital outputs, four operational amplifier, and one on-chip temperature sensor. The chip is housed in a small 28-pin SSOP package that, in many cases, makes it very convenient to position it right at the sensing site without using intermediate cables or wires.

5.6 Ratiometric Circuits

A powerful method for improving the accuracy of a sensor is a ratiometric technique, which is one of the most popular methods of signal conditioning. It should be emphasized, however, that the method is useful only if a source of error has a *multiplicative* nature but not additive; that is, the technique would be useless for reducing, for instance, thermal noise. On the other hand, it is quite potent for solving such problems as the dependence of a sensor's sensitivity to such factors as power-supply instability, ambient temperature, humidity, pressure, effects of aging, and so forth. The technique essentially requires the use of two sensors, of which one is the acting sensor, which responds to an external stimulus, and the other is a compensating sensor, which is either shielded from that stimulus or is insensitive to it. Both sensors must be exposed to all other external effects which may multiplicatively change their performance. The second sensor, which is often called *reference*, must be subjected to a reference stimulus, which is ultimately stable during the lifetime of the product. In many practical systems, the reference sensor must not necessarily be exactly similar to the acting sensor; however, its physical properties, which are subject to instabilities, should be the same. For example, Fig. 5.34A shows a simple temperature detector, where the acting sensor is a negative temperature coefficient (NTC) thermistor R_T. A stable reference resistor R_0 has a value equal to the resistance of the thermistor at some reference temperature, (e.g., at 25°C). Both are connected via an analog multiplexer to an amplifier with a feedback resistor R. The output signals produced by the sensor and the reference resistor respectively are

$$V_N = -\frac{ER}{R_T}, \qquad (5.39)$$

$$V_D = -\frac{ER}{R_0}.$$

It is seen that both voltages are functions of a power-supply voltage E and the circuit gain, which is defined by resistor R. That resistor as well as the power supply may be the sources of error. If two output voltages are fed into a divider circuit, the resulting signal may be expressed as $V_0 = kV_N/V_D = kR_0/R_T$, where k is the divider's gain. The output signal is not subject to either power-supply voltage or the amplifier gain. It depends only on the sensor and its reference resistor. This is true only if spurious variables (like the power supply or amplifier's gain) do not change rapidly; that is, they must not change appreciably within the multiplexing period. This requirement determines the rate of multiplexing.

A ratiometric technique essentially requires the use of a division. It can be performed by two standard methods: digital and analog. In a digital form, output signals from both the acting and the reference sensors are multiplexed and converted into binary codes in an A/D converter. Subsequently, a computer or a microprocessor performs the operation of a division. In an analog form, a divider may be a part of a signal conditioner or the interface circuit. A "divider" (Fig. 5.35A) produces an output

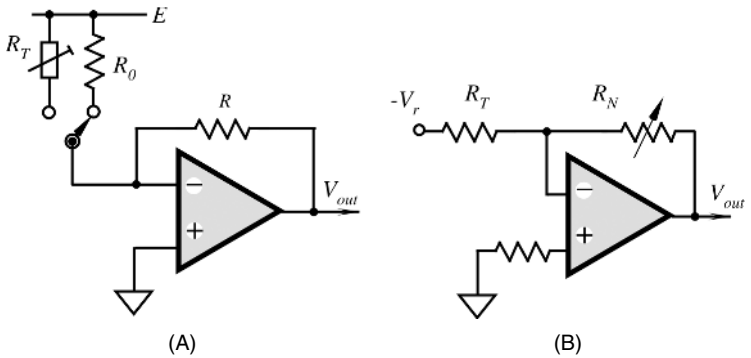


Fig. 5.34. Ratiometric temperature detector (A) and analog divider of resistive values (B).

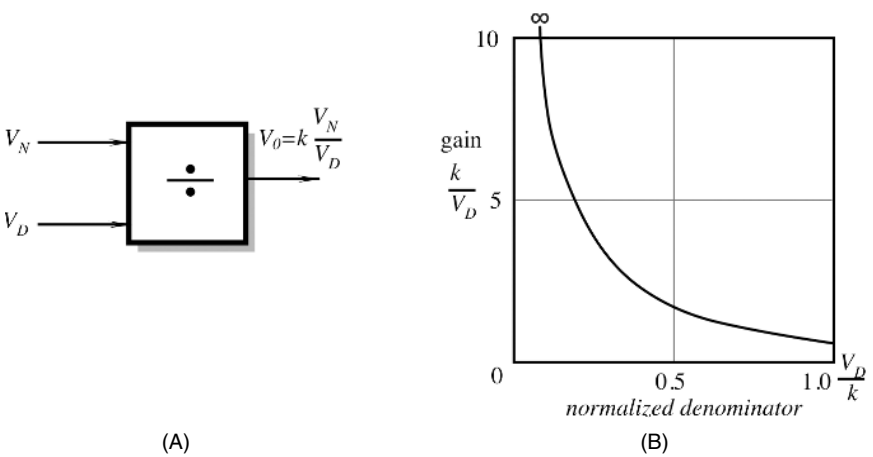


Fig. 5.35. Schematics of a divider (A) and gain of a divider as a function of a denominator (B).

voltage or current proportional to the ratio of two input voltages or currents:

$$V_0 = k\frac{V_N}{V_D}, \tag{5.40}$$

where the numerator is denoted as V_N, the denominator is V_D and k is equal to the output voltage, when $V_N = V_D$. The operating ranges of the variables (quadrants of operation) is defined by the polarity and magnitude ranges of the numerator and denominator inputs and of the output. For instance, if V_N and V_D are both either positive or negative, the divider is of a one-quadrant type. If the numerator is bipolar, the divider is a two-quadrant type. Generally, the denominator is restricted to a single polarity, because the transition from one polarity to another would require the denominator to pass through zero, which would call for an infinite output (unless

the numerator is also zero). In practice, the denominator is a signal from a reference sensor, which usually is of a constant value.

Division has long been the most difficult of the four arithmetic functions to implement with analog circuits. This difficulty stems primarily from the nature of division: the magnitude of a ratio becomes quite large, approaching infinity, for a denominator that is approaching zero (and a nonzero numerator). Thus, an ideal divider must have a potentially infinite gain and infinite dynamic range. For a real divider, both of these factors are limited by the magnification of drift and noise at low values of V_D; that is, the gain of a divider for a numerator is inversely dependent on the value of the denominator (Fig. 5.35B). Thus, the overall error is the net effect of several factors, such as gain dependence of denominator, numerator and denominator input errors, like offsets, noise, and drift (which must be much smaller than the smallest values of the input signals). In addition, the output of the divider must be constant for constant ratios of numerator and denominator, independent of their magnitudes; for example, $10/10 = 0.01/0.01 = 1$ and $1/10 = 0.001/0.01 = 0.1$.

5.7 Bridge Circuits

The Wheatstone bridge circuits are popular and very effective implementations of the ratiometric technique or a division technique on a sensor level. A basic circuit is shown in Fig. 5.36. Impedances Z may be either active or reactive; that is, they may be either simple resistances, as in piezoresistive gauges, or capacitors, or inductors. For the resistor, the impedance is R; for the ideal capacitor, the magnitude of its impedance is equal to $1/2\pi fC$; and for the inductor, it is $2\pi fL$, where f is the frequency of the current passing through the element. The bridge output voltage is represented by

$$V_{out} = \left(\frac{Z_1}{Z_1 + Z_2} - \frac{Z_3}{Z_3 + Z_4} \right) V_{\text{ref}}, \tag{5.41}$$

The bridge is considered to be in a balanced state when the following condition is met:

$$\frac{Z_1}{Z_2} = \frac{Z_3}{Z_4}. \tag{5.42}$$

Under the balanced condition, the output voltage is zero. When at least one impedance changes, the bridge becomes unbalanced and the output voltage goes either in a positive or negative direction, depending on the direction of the impedance change. To

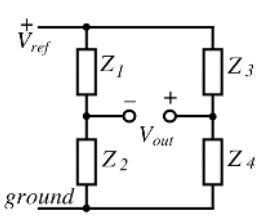


Fig. 5.36. General circuit of a Wheatstone bridge.

determine the bridge sensitivity with respect to each impedance (calibration constant), partial derivatives may be obtained from Eq. (5.41):

$$\begin{aligned}
\frac{\partial V_{\text{out}}}{\partial Z_1} &= \frac{Z_2}{(Z_1+Z_2)^2} V_{\text{ref}}, \\
\frac{\partial V_{\text{out}}}{\partial Z_2} &= -\frac{Z_1}{(Z_1+Z_2)^2} V_{\text{ref}}, \\
\frac{\partial V_{\text{out}}}{\partial Z_3} &= -\frac{Z_4}{(Z_3+Z_4)^2} V_{\text{ref}}, \\
\frac{\partial V_{\text{out}}}{\partial Z_4} &= \frac{Z_3}{(Z_3+Z_4)^2} V_{\text{ref}}.
\end{aligned} \tag{5.43}$$

By summing these equations, we obtain the bridge sensitivity:

$$\frac{\delta V_{\text{out}}}{V_{\text{ref}}} = \frac{Z_2 \delta Z_1 - Z_1 \delta Z_2}{(Z_1+Z_2)^2} - \frac{Z_4 \delta Z_3 - Z_3 \delta Z_4}{(Z_3+Z_4)^2}. \tag{5.44}$$

A closer examination of Eq. (5.44) shows that only the adjacent pairs of impedances (i.e., Z_1 and Z_2, Z_3 and Z_4) have to be identical in order to achieve the ratiometric compensation (such as the temperature stability, drift, etc.). It should be noted that impedances in the balanced bridge do not have to be equal, as long as a balance of the ratio (5.42) is satisfied. In many practical circuits, only one impedance is used as a sensor; thus for Z_1 as a sensor, the bridge sensitivity becomes

$$\frac{\delta V_{\text{out}}}{V_{\text{ref}}} = \frac{\delta Z_1}{4 Z_1}. \tag{5.45}$$

The resistive bridge circuits are commonly used with strain gauges, piezoresistive pressure transducers, thermistor thermometers, and other sensors when immunity against environmental factors is required. Similar arrangements are used with the capacitive and magnetic sensors for measuring force, displacement, moisture, and so forth.

5.7.1 Disbalanced Bridge

A basic Wheatstone bridge circuit (Fig. 5.37A) generally operates with a disbalanced bridge. This is called the *deflection* method of measurement. It is based on detecting the voltage across the bridge diagonal. The bridge output voltage is a nonlinear function of a disbalance Δ; however, for small changes ($\Delta < 0.05$), which often is the case, it may be considered quasilinear. The bridge maximum sensitivity is obtained when $R_1 = R_2$ and $R_3 = R$. When $R_1 \gg R_2$ or $R_2 \gg R_1$, the bridge output voltage is decreased. Assuming that $k = R_1/R_2$, the bridge sensitivity may be expressed as

$$\alpha = \frac{V}{R} \frac{k}{(k+1)^2}. \tag{5.46}$$

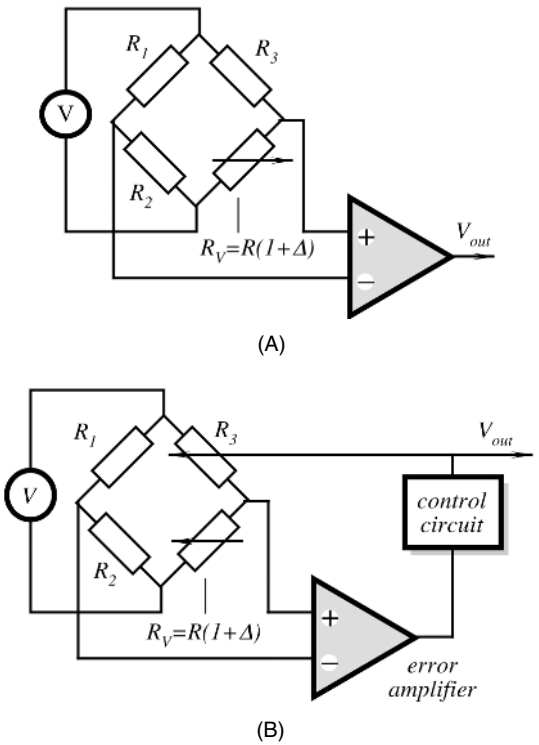


Fig. 5.37. Two methods of using a bridge circuit: (A) disbalanced bridge and (B) balanced bridge with a feedback control.

A normalized graph calculated according to this equation is shown in Fig. 5.38. It indicates that the maximum sensitivity is achieved at $k = 1$. However, the sensitivity drops relatively little for the range where $0.5 < k < 2$. If the bridge is fed by a current source, rather than by a voltage source, its output voltage for small Δ and a single-variable component is represented by

$$V_{\text{out}} \approx I \frac{k\Delta}{2(k+1)}, \tag{5.47}$$

where I is the excitation current.

5.7.2 Null-Balanced Bridge

Another method for using a bridge circuit is called a *null-balance*. The method overcomes the limitation of small changes (Δ) in the bridge arm to achieve a good linearity. The null-balance essentially requires that the bridge *always* be maintained at the balanced state. To satisfy the requirement for a bridge balance (5.42), another arm of the bridge should vary along with the arm used as a sensor. Figure 5.37B illustrates

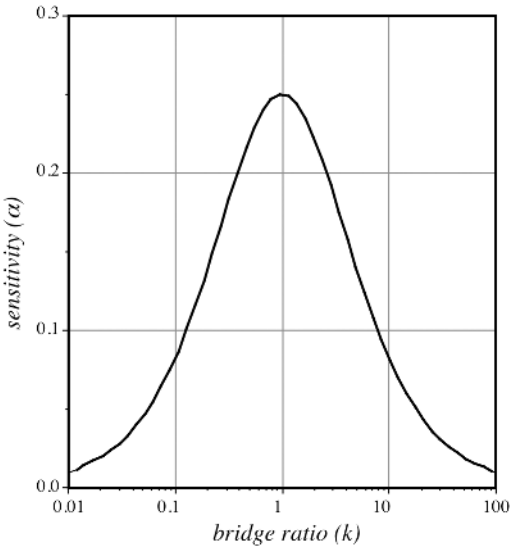


Fig. 5.38. Sensitivity of a disbalanced bridge as a function of impedance ratio.

this concept. A control circuit modifies the value of R_3 on the command from the error amplifier. The output voltage may be obtained from the control signal of the balancing arm R_3. For example, both R_v and R_3 may be photoresistors. The R_3 photoresistor could be interfaced with a light-emitting diode (LED) which is controlled by the error amplifier. Current through the LED becomes a measure of resistance R_v, and, subsequently, of the light intensity detected by the sensor.

5.7.3 Temperature Compensation of Resistive Bridge

The connection of four resistive components in a Wheatstone bridge configuration is used quite extensively in measurements of temperature, force, pressure, magnetic fields, and so forth. In many of these applications, sensing resistors exhibit temperature sensitivity. This results in the temperature sensitivity of a transfer function, which, by using a linear approximation, may be expressed by Eq. (2.1) of Chapter 2. In any detector, except that intended for temperature measurements, this temperature dependence has a highly undesirable effect, which usually must be compensated for. One way to do a compensation is to couple a detector with a temperature-sensitive device, which can generate a temperature-related signal for the hardware or software correction. Another way to do a temperature compensation is to incorporate it directly into the bridge circuit. Let us analyze the Wheatstone bridge output signal with respect to its excitation signal V_e. We consider all four arms in the bridge being responsive to a stimulus with the sensitivity coefficient, α, so that each resistor has the value

$$R_i = R(1 \pm \alpha s), \tag{5.48}$$

where R is the nominal resistance and s is the stimulus (e.g., pressure or force), and the sensitivity, α, is defined as

$$\alpha = \frac{1}{R}\frac{dR}{ds}. \tag{5.49}$$

An output voltage from the bridge is

$$V_{\text{out}} = V_e \alpha s + V_0, \tag{5.50}$$

where V_0 is the offset voltage resulting from the initial bridge imbalance. If the bridge is not properly balanced, the offset voltage may be a source of error. However, an appropriate trimming of the bridge sensor during either its fabrication or in application apparatus may reduce this error to an acceptable level. In any event, even if the offset voltage is not properly compensated for, its temperature variations usually are several orders of magnitude smaller than that of the sensor's transfer function. In this discussion, we consider V_0 temperature independent ($dV_0/dT = 0$); however, for a broad temperature range (wider than $\pm 15°\text{C}$) V_0 should not be discounted.

In Eq. (5.48), sensitivity α generally is temperature dependent for many sensors and is a major source of inaccuracy. It follows from Eq. (5.49) that α may vary if R is temperature dependent or when dR/ds is temperature dependent. If the bridge has a positive temperature coefficient of resistivity, the coefficient α decreases with temperature, or, it is said, it has a negative TCS (temperature coefficient of sensitivity). Taking a partial derivative with respect to temperature T, from Eq. (5.50) we arrive at

$$\frac{\partial V_{\text{out}}}{\partial T} = s\left(\alpha\frac{\partial V_e}{\partial T} + \frac{\partial \alpha}{\partial T}V_e\right). \tag{5.51}$$

A solution of this equation is the case when the output signal does not vary with temperature: $\partial V_{\text{out}}/\partial T = 0$. Then, the following holds:

$$\alpha\frac{\partial V_e}{\partial T} = -\frac{\partial \alpha}{\partial T}V_e, \tag{5.52}$$

and, finally,

$$\frac{1}{V_e}\frac{\partial V_e}{\partial T} = -\frac{1}{\alpha}\frac{\partial \alpha}{\partial T} = -\beta, \tag{5.53}$$

where β is the TCS of the bridge arm.

The above is a condition for an *ideal* temperature compensation of a fully symmetrical Wheatstone bridge; that is, to compensate for temperature variations in α, the excitation voltage, V_e, must change with temperature at the same rate and with opposite sign. To control V_e, several circuits were proven to be useful [9]. Figure 5.39 shows a general circuit which incorporates a temperature compensation network to control voltage V_e across the bridge according to a predetermined function of temperature. Several options of the temperature compensation network are possible.

Option 1: Use of a temperature sensor as a part of the compensation network. Such a network may be represented by an equivalent impedance R_t, and an entire bridge can be represented by its equivalent resistance R_B. Then, the voltage across the bridge is

$$V_e = E\frac{R_B}{R_B + R_t}. \tag{5.54}$$

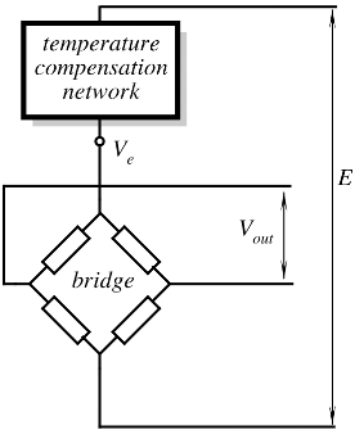


Fig. 5.39. General circuit of a bridge temperature compensation.

Taking the derivative with respect to temperature, we get

$$\frac{\partial V_e}{\partial T} = E\left[\frac{1}{R_B + R_t}\frac{\partial R_B}{\partial T} - \frac{R_B}{(R_B + R_t)^2}\left(\frac{\partial R_B}{\partial T} + \frac{\partial R_t}{\partial T}\right)\right], \tag{5.55}$$

and substituting Eq. (5.54) into Eq. (5.55), we arrive at the compensation condition:

$$\frac{1}{V_e}\frac{\partial V_e}{\partial T} = \frac{1}{R_B}\frac{\partial R_B}{\partial T} - \frac{1}{R_B + R_t}\left(\frac{\partial R_B}{\partial T} + \frac{\partial R_t}{\partial T}\right). \tag{5.56}$$

Because for a bridge with four sensitive arms, $R_B = R$, and $(1/R)(\partial R/\partial T) = \gamma$ which is a temperature coefficient of bridge arm resistance, R (TCR), Eq. (5.56) according to Eq. (5.53) must be equal to a negative TCS:

$$-\beta = \gamma - \frac{1}{R + R_t}\left(\frac{\partial R}{\partial T} + \frac{\partial R_t}{\partial T}\right). \tag{5.57}$$

This states that such a compensation is useful over a broad range of excitation voltages because E is not a part of the equation.[5] To make it work, the resistive network R_t must incorporate a temperature-sensitive resistor, (e.g., a thermistor). When R, β and $\partial R/\partial T$ are known, Eq. (5.57) can be solved to select R_t. This method requires a trimming of the compensating network to compensate not only for TCS and TCR, but for V_e as well. The method, although somewhat complex, allows for a broad range of temperature compensation from -20 to $+70$°C and with somewhat reduced performance or with a more complex compensating network, from -40 to 100°C. Figure 5.40A shows an example of the compensating network which incorporates an NTC thermistor $R°$ and several trimming resistors. An example of such a compensation is a Motorola pressure sensor, PMX2010, which contains a diffused into silicon temperature-compensating resistors which calibrate offset and compensate for temperature variations. The resistors are laser trimmed on-chip during the calibrating process to assure high stability over a broad temperature range.

[5] This demands that the compensation network contain no active components, such as diodes, transistors, and so forth.

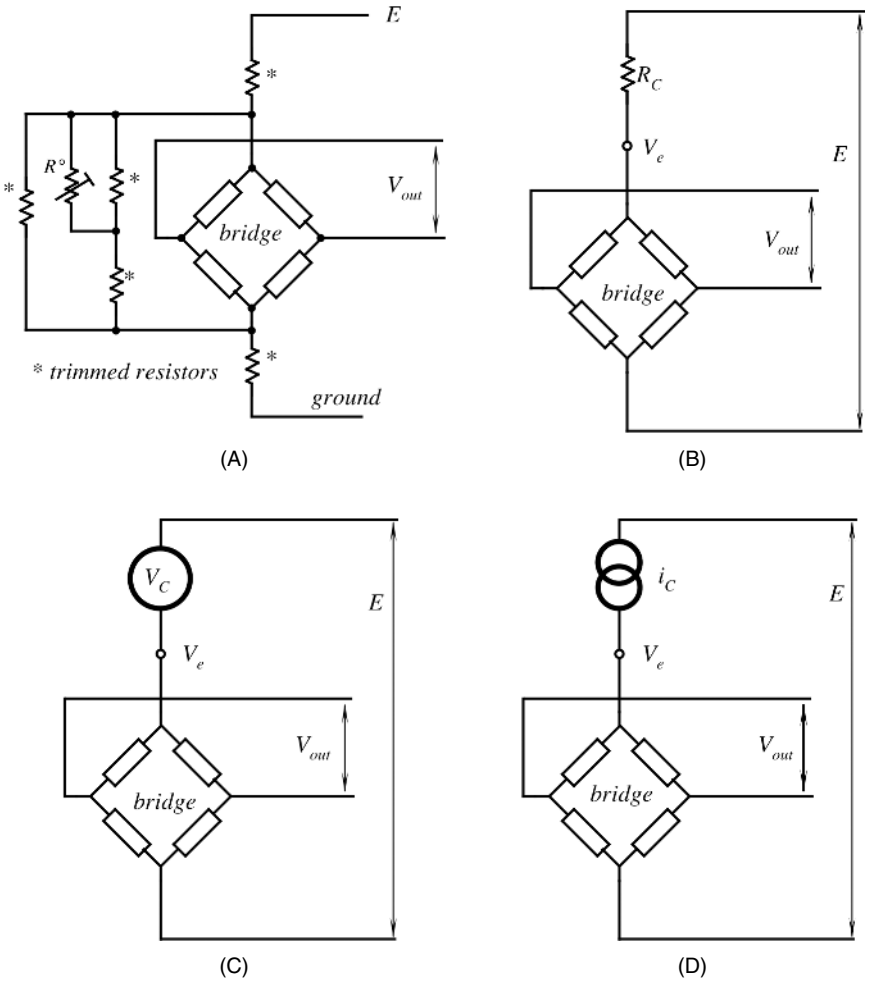


Fig. 5.40. Temperature compensation of a bridge circuit: (A) with NTC thermistor; (B) with a fixed resistor; (C) with a temperature-controlled voltage source; (D) with a current source.

Option 2: The compensation network is a fixed resistor. This is the most popular temperature compensation of a resistive Wheatstone bridge. A fixed resistor (Fig. 5.40B) $R_t = R_c$ must have low-temperature sensitivity (50 ppm or less). It can be stated that

$$\frac{1}{R_c}\frac{\partial R_c}{\partial T} = 0, \tag{5.58}$$

and Eq. (5.57) is simplified to

$$-\beta = \frac{\partial R}{\partial T}\left(\frac{1}{R} - \frac{1}{R + R_c}\right). \tag{5.59}$$

It can be solved for the temperature-compensating resistor

$$R_c = -\frac{\beta R}{\partial R/\partial T + \beta R}. \tag{5.60}$$

Then, the compensating resistor is

$$R_c = -R\frac{\beta}{\gamma + \beta}. \tag{5.61}$$

The minus sign indicates that the equation is true for negative TCS β. Thus, when TCR, TCS, and a nominal resistance of the bridge arm are known, a simple stable resistor in series with voltage excitation E can provide a quite satisfactory compensation. It should be noted, however, that according to Eq. (5.59), to use this method, the TCS of the bridge arm must be smaller than its TCR ($|\beta| < \gamma$). As for Option 1, this circuit is ratiometric with respect to the operating voltage, E, meaning that the compensation works over a broad range of power-supply voltages. Selecting R_c according to Eq. (5.61) may result in a very large compensating resistance, which may be quite inconvenient in many applications. The sensor's TCR can be effectively reduced by adding a resistor in parallel with the bridge. When a large resistor R_c is used, this method of compensation becomes similar to Option 4 because a large resistor operates as a "quasi"-current source.

A first impression of this option is that it seems like a perfect solution—just one resistor ideally compensates for a temperature drift. However, this option does not yield satisfactory results for broad temperature ranges or precision applications. For instance, to select an appropriate R_c, γ, and β must be precisely known; that is, each actual bridge must be characterized, which is not acceptable in low-cost applications. Using typical rather than actual values may result in span errors on the order of 100 ppm/°C. Further, large resistors R_c cause lower output voltages and a reduced signal-to-noise ratio. Practically, the usefulness of this compensation is limited to the range 25 ± 15°C.

Option 3: A compensating network contains a temperature-controlled voltage source, (e.g., a diode or transistor) (Fig. 5.40C). For this circuit, to satisfy a condition for the best compensation of TCS β, the temperature sensitivity β_c of the voltage source V_c must be

$$\beta_c = \beta\left(\frac{E}{V_c} - 1\right). \tag{5.62}$$

Because β is a parameter of the bridge, by manipulating E and V_c, one can select the optimum compensation. Because the compensating circuit contains a voltage source, it is not ratiometric with respect to the power supply. For the operation, this option requires a regulated source of E. An obvious advantage of the circuit is simplicity because diodes and transistors with predicable temperature characteristics are readily available. An obvious disadvantage of this method is a need to operate the sensor at a fixed specified voltage. A useful temperature range for this method is about 25 ± 25°C.

Option 4: A current source is employed as an excitation circuit (Fig. 5.40D). This circuit requires that the bridge possesses a particular property. Its TCR (β) must be equal to TCS (α) with the opposite sign:

$$\alpha = -\beta. \tag{5.63}$$

The voltage across the bridge is equal to

$$V_e = i_c R_B. \tag{5.64}$$

Because the current source is temperature independent and, for a bridge with four identical arms, $R_B = R$, then

$$\frac{\partial V_e}{\partial T} = I_c \frac{\partial R}{\partial T}, \tag{5.65}$$

and dividing Eq. (5.65) by Eq. (5.64) we arrive at

$$\frac{1}{V_c}\frac{\partial V_e}{\partial T} = \frac{1}{R}\frac{\partial R}{\partial T}. \tag{5.66}$$

If condition (5.63) is fulfilled, we receive a provision of an ideal compensation as defined by Eq. (5.52). Unfortunately, this method of compensation has a limitation similar to that for Option 2—specifically, a reduced output voltage and a need for individual sensor characterization if used in a broad temperature range. Nevertheless, this method is acceptable when the accuracy of 1–2% of FS over 50°C is acceptable.

The above options provide a framework of the compensating techniques. While designing a practical circuit, many variables must be accounted for: temperature range, allowable temperature error, environmental conditions, size, cost, and so forth. Therefore, we cannot recommend a universal solution; the choice of the most appropriate option must be a result of a typical engineering compromise.

5.7.4 Bridge Amplifiers

The bridge amplifiers for resistive sensors are probably the most frequently used sensor interface circuits. They may be of several configurations, depending on the required bridge grounding and availability of either grounded or floating reference voltages. Figure 5.41A shows the so-called active bridge, where a variable resistor (the sensor) is floating (i.e., isolated from ground) and is connected into a feedback of the OPAM. If a resistive sensor can be modeled by a first-order function

$$R_x \approx R_0(1+\alpha), \tag{5.67}$$

then, a transfer function of this circuit is

$$V_{\text{out}} = -\frac{1}{2}\alpha V. \tag{5.68}$$

A circuit with a floating bridge and floating reference voltage source V is shown in Fig. 5.41B. This circuit may provide gain which is determined by a feedback resistor whose value is nR_0:

$$V_{\text{out}} = (1+n)\alpha\frac{V}{4}\frac{1}{1+\alpha/2} \approx (1+n)\alpha\frac{V}{4}. \tag{5.69}$$

A bridge with the asymmetrical resistors ($R \neq R_0$) may be used with the circuit shown in Fig. 5.41C. It requires a floating reference voltage source V:

$$V_{\text{out}} = n\alpha\frac{V}{4}\frac{1}{1+\alpha/2} \approx n\alpha\frac{V}{4}. \tag{5.70}$$

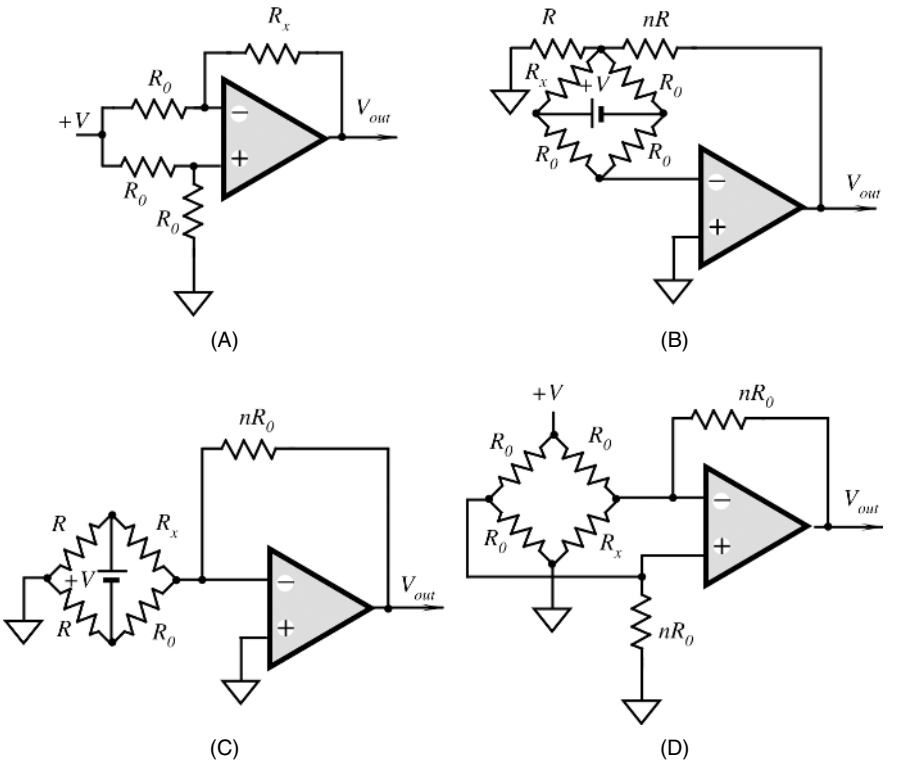


Fig. 5.41. Connection of operational amplifiers to resistive bridge circuits (disbalanced mode).

When a resistive sensor is grounded and a gain from the interface circuit is desirable, the schematic shown in Fig. 5.41D may be employed. Its transfer function is determined from

$$V_{\text{out}} = -\frac{n}{2}\frac{V}{1+1/2n}\frac{\alpha}{1+\alpha} \approx -\frac{n}{2}\frac{V}{1+1/2n}\alpha. \tag{5.71}$$

When the bridge is perfectly balanced, the output voltage V_{out} is equal to one-half of the bridge excitation voltage $+V$. To better utilize the operational amplifier open-loop gain, the value of n should not exceed 50.

5.8 Data Transmission

A signal from a sensor may be transmitted to the receiving end of the system either in a digital format or analog. In most cases, a digital format essentially requires the use of an analog-to-digital converter at the sensor's site. The transmission in a digital format has several advantages, the most important of which is noise immunity. The

transmission of digital information is beyond the scope of this book; thus, we will not discuss it further. In many cases, however, digital transmission can not be done for several reasons. Then, the sensor signals are transmitted to the receiving site in an analog form. Depending on connection, they can be divided into a two-, four-, and six-wire methods.

5.8.1 Two-Wire Transmission

Two-wire analog transmitters are used to couple sensors to control and monitoring devices in the process industry [10]. When, for example, a temperature measurements is taken within a process, a two-wire transmitter relays that measurement to the control room or interfaces the measurement directly to a process controller. Two wires can be used to transmit either voltage or current; however, current was accepted as an industry standard. It varies in the range 4–20 mA, which represents an entire span of the input stimulus. Zero stimulus corresponds to 4 mA while the maximum is at 20mA. There are two advantages of using current rather than voltage, as is illustrated in Fig. 5.42. Two wires join the controller site with the sensor site. At the sensor site, there is a sensor which is connected to the so-called *two-wire transmitter*. The transmitter may be a voltage-to-current converter; that is, it converts the sensor signal into a variable current. At the controller site, there is a voltage source that can deliver current up to 20 mA. The two wires form a current loop, which, at the sensor's side, has the sensor and a transmitter, whereas at the controller side, it has a load resistor and a power supply which are connected in series. When the sensor signal varies, the transmitter's output resistance varies accordingly, thus modulating the current in the range between 4 and 20 mA. The same current which carries information is also used by the transmitter and the sensor to provide their operating power. Obviously, even for the lowest output signal which produces 4 mA current, that 4 mA must be sufficient to power the transmitting side of the loop. The loop current causes a voltage drop across the load resistor at the controller side. This voltage is a received signal which is suitable for further processing by the electronic circuits. An advantage of the two-

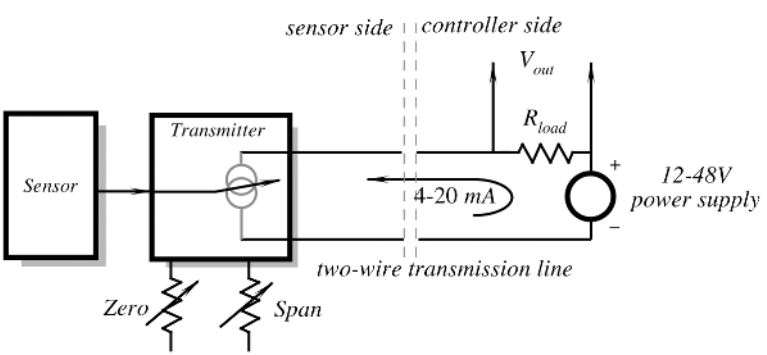


Fig. 5.42. Two-wire 20-mA analog data transmission.

wire method is that the transmitting current is independent of the connecting wires' resistance and thus of the transmission line length (obviously, within the limits).

5.8.2 Four-Wire Sensing

Sometimes, it is desirable to connect a resistive sensor to a remotely located interface circuit. When such a sensor has a relatively low resistance (e.g., it is normal for the piezoresistors or RTDs to have resistances in the order of 100Ω), connecting wire resistances pose a serious problem because they alter the excitation voltage across the bridge. The problem can be solved by using the so-called *four-wire method* (Fig. 5.43A). It allows measuring the resistance of a remote resistor without measuring the resistances of the connecting conductors. A resistor which is the subject of measurement is connected to the interface circuit through four rather than two wires. Two wires are connected to a current source and two others to the voltmeter. A constant-current source (current pump) has a very high output resistance; therefore, the current which it pushes through the loop is almost independent of any resistances r in that loop. An input impedance of a voltmeter is very high; hence, no current is diverted from the current loop to the voltmeter. The voltage drop across the resistor R_x is

$$V_x = R_x i_0, \tag{5.72}$$

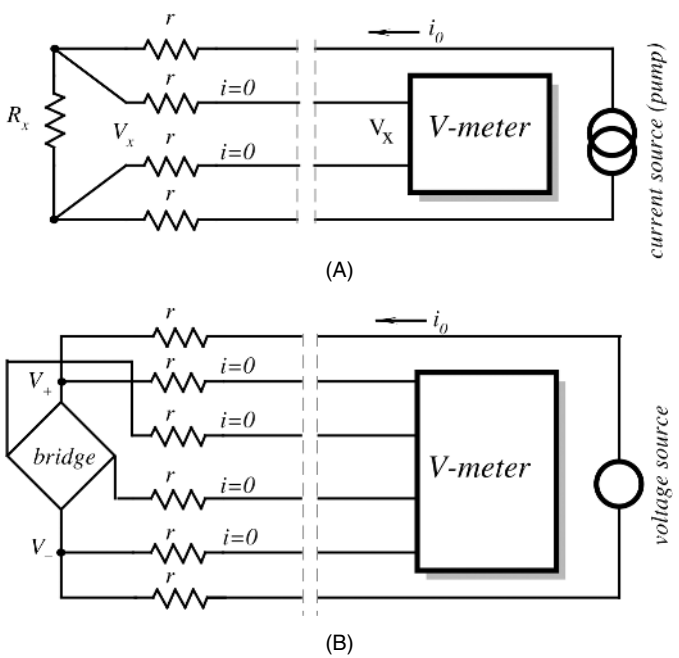


Fig. 5.43. Remote measurements of resistances: (A) four-wire method; (B) six-wire measurement of a bridge.

which is independent of any resistances r of the connecting wires. The four-wire method is a very powerful means of measuring the resistances of remote detectors and is used in industry and science quite extensively.

5.8.3 Six-Wire Sensing

When a Wheatstone bridge circuit is remotely located, voltage across the bridge plays an important role in the bridge temperature stability, as was shown in Section 5.7. That voltage often should be either measured or controlled. Long transmitting wires may introduce unacceptably high resistance in series with the bridge excitation voltage, which interferes with the temperature compensation. The problem may be solved by providing two additional wires to feed the bridge with voltage and to dedicate two wires to measuring the voltage across the bridge (Fig. 5.43B). The actual excitation voltage across the bridge and the bridge differential output voltage are measured by a high-input impedance voltmeter with negligibly small input currents. Thus, the accurate bridge voltages are available at the data processing site without being affected by the long transmission lines.

5.9 Noise in Sensors and Circuits

Noise in sensors and circuits may present a substantial source of errors and should be seriously considered. "Like diseases, noise is never eliminated, just prevented, cured, or endured, depending on its nature, seriousness, and the cost/difficulty of treating" [11]. There are two basic classifications of noise for a given circuit: *inherent* noise, which is noise arising within the circuit, and *interference (transmitted)* noise, which is noise picked up from outside the circuit.

Any sensor, no matter how well it was designed, never produces an electric signal which is an ideal representation of the input stimulus. Often, it is a matter of judgment to define the goodness of the signal. The criteria for this are based on the specific requirements to accuracy and reliability. Distortions of the output signal can be either systematic or stochastic. The former are related to the sensor's transfer function, its linearity, dynamic characteristics, and so forth. All are the result of the sensor's design, manufacturing tolerances, material quality, and calibration. During a reasonably short time, these factors either do not change or drift relatively slowly. They can be well defined, characterized, and specified (see Chapter 2). In many applications, such a determination may be used as a factor in the error budget and can be taken into account. Stochastic disturbances, on the other hand, often are irregular, unpredictable to some degree, and may change rapidly. Generally, they are termed *noise*, regardless of their nature and statistical properties. It should be noted that the word *noise*, in association with audio equipment noise, is often mistaken for an irregular, somewhat fast-changing signal. We use this word in a much broader sense for all disturbances, either in stimuli, environment, or components of sensors and circuits from dc to the upper operating frequencies.

5.9.1 Inherent Noise

A signal which is amplified and converted from a sensor into a digital form should be regarded not just by its magnitude and spectral characteristics but also in terms of a digital resolution. When a conversion system employs an increased digital resolution, the value of the least significant bit (LSB) decreases. For example, the LSB of a 10-bit system with a 5-V full scale is about 5 mV; the LSB of 16 bits is 77 μV. This by itself poses a significant problem. It makes no sense to employ, say, a 16-bit resolution system, if combined noise is, for example, 300 μV. In the real world, the situation is usually much worse. There are almost no sensors which are capable of producing a 5-V full-scale output signals. Most of them require amplification. For instance, if a sensor produces a full-scale output of 5 mV, at a 16-bit conversion it would correspond to a LSB of 77 nV—an extremely small signal which makes amplification an enormous task by itself. Whenever a high resolution of a conversion is required, all sources of noise must be seriously considered. In the circuits, noise can be produced by the monolithic amplifiers and other components which are required for the feedback, biasing, bandwidth limiting, and so forth.

Input offset voltages and bias currents may drift. In dc circuits, they are indistinguishable from low-magnitude signals produced by a sensor. These drifts are usually slow (within a bandwidth of tenths and hundredths of a hertz); therefore, they are often called ultralow-frequency noise. They are equivalent to randomly (or predictable—say, with temperature) changing voltage and current offsets and biases. To distinguish them from the higher-frequency noise, the equivalent circuit (Fig. 5.3) contains two additional generators. One is a *voltage offset* generator e_0 and the other is a *current bias* generator i_0 . The noise signals (voltage and current) result from physical mechanisms within the resistors and semiconductors that are used to fabricate the circuits. There are several sources of noise whose combined effect is represented by the noise voltage and current generators.

One cause for noise is a discrete nature of electric current because current flow is made up of moving charges, and each charge carrier transports a definite value of charge (the charge of an electron is 1.6×10^{-19} C). At the atomic level, current flow is very erratic. The motion of the current carriers resembles popcorn popping. This was chosen as a good analogy for current flow and has nothing to do with the "popcorn noise," which we will discuss later. As popcorn, the electron movement may be described in statistical terms. Therefore, one never can be sure about very minute details of current flow. The movement of carriers are temperature related and noise power, in turn, is also temperature related. In a resistor, these thermal motions cause Johnson noise to result [12]. The mean-square value of noise voltage (which is representative of noise power) can be calculated from

$$\overline{e_n^2} = 4kTR\Delta f \left(\frac{V^2}{\text{Hz}}\right), \tag{5.73}$$

where $k = 1.38 \times 10^{-23}$ J/K (Boltzmann constant), T is the temperature (in K), R is the resistance (in Ω), and Δf is the bandwidth over which the measure-

ment is made (in Hz). For practical purposes, noise density per $\sqrt{\text{Hz}}$ generated by a resistor at room temperature may be estimated from a simplified formula: $\overline{e_n} \approx 0.13\sqrt{R}$ in nV/$\sqrt{\text{Hz}}$. For example, if the noise bandwidth is 100 Hz and the resistance of concern is 10 MΩ ($10^7\,\Omega$), the average noise voltage is estimated as $\overline{e_n} \approx 0.13\sqrt{10^7}\sqrt{100} = 4{,}111\text{nV} \approx 4\mu\text{V}$.

Even a simple resistor is a source of noise. It behaves as a perpetual generator of electric signal. Naturally, relatively small resistors generate extremely small noise; however, in some sensors, Johnson noise must be taken into account. For instance, a pyroelectric detector uses a bias resistor on the order of 50 GΩ. If a sensor is used at room temperature within a bandwidth of 100 Hz, one may expect the average noise voltage across the resistor to be on the order of 0.3 mV—a very high value. To keep noise at bay, bandwidths of the interface circuits must be maintained small, just wide enough to pass the minimum required signal. It should be noted that noise voltage is proportional to the square root of the bandwidth. It implies that if we reduce the bandwidth 100 times, the noise voltage will be reduced by a factor of 10. The Johnson noise magnitude is constant over a broad range of frequencies. Hence, it is often called *white noise* because of the similarity to white light, which is composed of all the frequencies in the visible spectrum.

Another type of noise results because of dc current flow in semiconductors. It is called *shot noise*; the name was suggested by Schottky not in association with his own name but rather because this noise sounded like "a hail of shot striking the target" nevertheless, shot noise is often called *Schottky noise*. Shot noise is also white noise. Its value becomes higher with the increase in the bias current. This is the reason why in FET and CMOS semiconductors current noise is quite small. For a bias current of 50 pA, it is equal to about 4 fA/$\sqrt{\text{Hz}}$—an extremely small current which is equivalent to the movement of about 6000 electrons per second. A convenient equation for shot noise is

$$i_{sn} = 5.7 \times 10^{-4}\sqrt{I\Delta f}, \tag{5.74}$$

where I is a semiconductor junction current in picoamperes and Δf is a bandwidth of interest in hertz.

An additional ac noise mechanism exists at low frequencies (Fig. 5.44). Both the noise voltage and noise current sources have a spectral density roughly proportional to $1/f$, which is called the *pink noise*, because of the higher noise contents at lower frequencies (lower frequencies are also on the red side of the visible spectrum). This

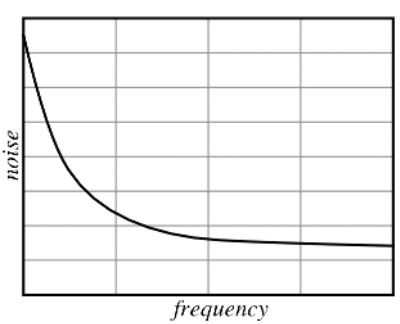


Fig. 5.44. Spectral distribution of $1/f$ "pink" noise.

$1/f$ noise occurs in all conductive materials; therefore, it is also associated with resistors. At extremely low frequencies, it is impossible to separate the $1/f$ noise from dc drift effects. The $1/f$ noise is sometimes called a flicker noise. Mostly, it is pronounced at frequencies below 100 Hz, where many sensors operate. It may dominate Johnson and Schottky noises and becomes a chief source of errors at these frequencies. The magnitude of pink noise depends on current passing through the resistive or semiconductive material. Currently, progress in semiconductor technology resulted in significant reduction of $1/f$ noise in semiconductors; however, when designing a circuit, it is a good engineering practice to use a metal film or wire-wound resistors in sensors and the front stages of interface circuits wherever significant currents flow through the resistor and low noise at low frequencies is a definite requirement.

A peculiar ac noise mechanism is sometimes seen on the screen of an oscilloscope when observing the output of an operational amplifier—a principal building block of many sensor interface circuits. It looks like a digital signal transmitted from outer space; noise has a shape of square pulses having variable duration of many milliseconds. This abrupt type of noise is called *popcorn noise* because of the sound it makes coming over a loudspeaker. Popcorn noise is caused by defects that are dependent on the integrated-circuit manufacturing techniques. Thanks to advances fabricating technologies, this type of noise is drastically reduced in modern semiconductor devices.

A combined noise from all voltage and current sources is given by the sum of squares of individual noise voltages:

$$e = \sqrt{e_{n1}^2 + e_{n2}^2 + \cdots + (R_1 i_{n1})^2 + (R_1 i_{n2})^2 + \cdots}. \tag{5.75}$$

A combined random noise may be presented by its root mean square (r.m.s) value, which is

$$E_{\text{rms}} = \sqrt{\frac{1}{T}\int_0^T e^2\,dt}, \tag{5.76}$$

where T is the time of observation, e is the noise voltage, and t is time.

Also, noise may be characterized in terms of the peak values which are the differences between the largest positive and negative peak excursions observed during an arbitrary interval. For some applications, in which peak-to-peak (p-p) noise may limit the overall performance (in a threshold-type devices), p-p measurement may be essential. Yet, due to a generally Gaussian distribution of noise signal, p-p magnitude is very difficult to measure in practice. Because r.m.s. values are so much easier to measure repeatedly and they are the most usual form for presenting noise data noncontroversially, Table 5.3 should be useful for estimating the probabilities of exceeding various peak values given by the r.m.s. values. The casually observed p-p noise varies between three times the r.m.s. and eight times the r.m.s., depending on the patience of observer and amount of data available.

5.9.2 Transmitted Noise

A large portion of environmental stability is attributed to the resistance of a sensor and an interface circuit to noise which originated in external sources. Figure 5.45 is

Table 5.3. Peak-to-Peak Value versus r.m.s. (for Gaussian Distribution)

Nominal p-p voltage	% of Time That Noise Will Exceed Nominal p-p Value
2 × r.m.s.	32.0
3 × r.m.s.	13.0
4 × r.m.s.	4.6
5 × r.m.s.	1.2
6 × r.m.s.	0.27
7 × r.m.s.	0.046
8 × r.m.s.	0.006

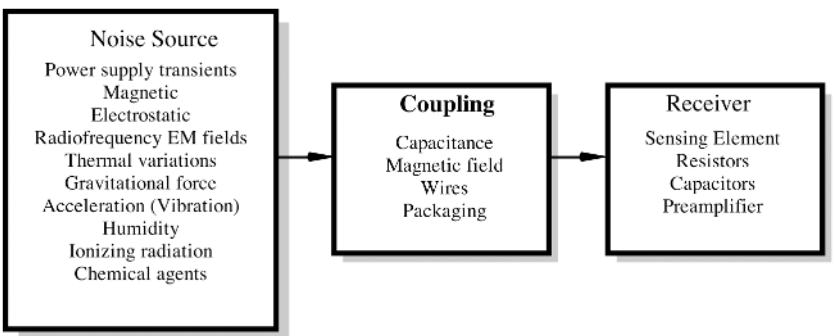


Fig. 5.45. Sources and coupling of transmitted noise.

a diagram of the transmitted noise propagation. Noise comes from a source which often can be identified. Examples of the sources are voltage surges in power lines, lightning, change in ambient temperature, sun activity, and so forth. These interferences propagate toward the sensor and the interface circuit, and eventually appear at the output. However, before that, they somehow must affect the sensing element inside the sensor, its output terminals, or the electronic components in a circuit. The sensor and the circuit function as receivers of the interferences.

There can be several classifications of transmitted noise, depending on how it affects the output signal, how it enters the sensor or circuit, and so forth. With respect to its relation to the output signals, noise can be either *additive* or *multiplicative*.

Additive noise e_n is added to the useful signal V_s and mixed with it as a fully independent voltage (or current):

$$V_{\text{out}} = V_s + e_n. \tag{5.77}$$

An example of such a disturbance is depicted in Fig. 5.46B. It can be seen that the noise magnitude does not change when the actual (useful) signal changes. As long as the sensor and interface electronics can be considered linear, the additive noise magnitude is totally independent of the signal magnitude, and if the signal is equal to zero, the output noise will be present still.

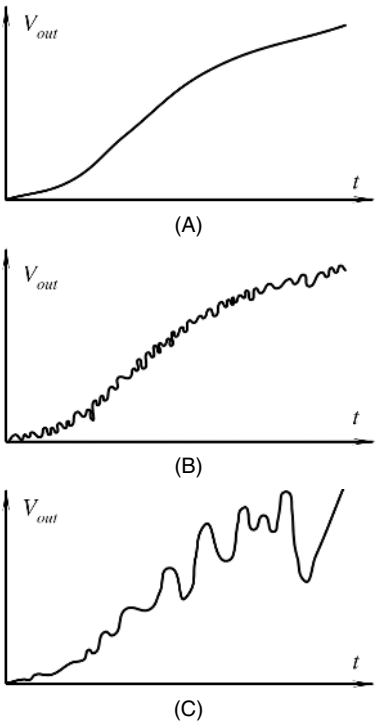


Fig. 5.46. Types of noise: (A) noise-free signal; (B) additive noise; (C) multiplicative noise.

Multiplicative noise affects the sensor's transfer function or the circuit's nonlinear components in such a manner as the V_s signal's value becomes altered or *modulated* by the noise:

$$V_{\text{out}} = [1 + N(t)]V_s, \tag{5.78}$$

where $N(t)$ is a function of noise. An example of such noise is shown in Fig. 5.46C. Multiplicative noise at the output disappears or becomes small (it also becomes additive) when the signal's magnitude nears zero. Multiplicative noise grows together with the signal's V_s magnitude. As its name implies, multiplicative noise is a result of multiplication (which is essentially a nonlinear operation) of two values where one is a useful signal and the other is a noise-dependent value.

To improve noise stability against transmitted additive noise, quite often sensors are combined in pairs; that is, they are fabricated in a dual form whose output signals are subtracted from one another (Fig. 5.47). This method is called a *differential* technique. One sensor of the pair (it is called the main sensor) is subjected to a stimulus of interest s_1, while the other (reference) is shielded from stimulus perception.

Since additive noise is specific for the linear or quasilinear sensors and circuits, the reference sensor does not have to be subjected to any particular stimulus. Often, it may be equal to zero. It is anticipated that both sensors are subjected to identical *transmitted* noise (noise generated inside the sensor cannot be canceled by a differential technique), which it is said is a common-mode noise. This means that noisy

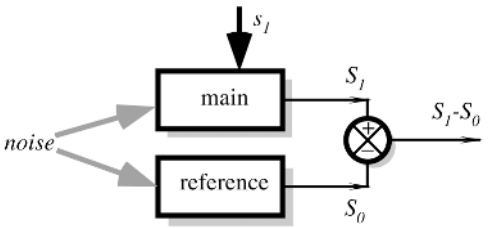


Fig. 5.47. Differential technique.

effects at both sensors are in phase and have the same magnitude. If both sensors are identically influenced by common-mode spurious stimuli, the subtraction removes the noise component. Such a sensor is often called either a dual or a *differential* sensor. The quality of noise rejection is described by a number called the *common-mode rejection ratio* (CMRR):

$$\text{CMRR} = 0.5\frac{S_1 + S_0}{S_1 - S_0}, \tag{5.79}$$

where S_1 and S_0 are output signals from the main and reference sensors, respectively. CMRR may depend on the magnitudes of stimuli and usually becomes smaller at greater input signals. The ratio shows how many times stronger the actual stimulus will be represented at the output, with respect to a common-mode noise having the same magnitude. The value of the CMRR is a measure of the sensor's symmetry. To be an effective means of noise reduction, both sensors must be positioned as close as possible to each other; they must be very identical and subjected to the same environmental conditions. Also, it is very important that the reference sensor be reliably shielded from the actual stimulus; otherwise, the combined differential response will be diminished.

To reduce transmitted multiplicative noise, a ratiometric technique is quite powerful (see Section 5.6 for the circuits' description). Its principle is quite simple. The sensor is fabricated in a dual form where one part is subjected to the stimulus of interest and both parts are subjected to the same environmental conditions, which may cause transmitted multiplicative noise. The second sensor is called *reference* because a constant environmentally stable reference stimulus s_0 is applied to its input. For example, the output voltage of a sensor in a narrow temperature range may be approximated by

$$V_1 \approx [1 + \alpha(T - T_0)] f(s_1), \tag{5.80}$$

where α is the temperature coefficient of the sensor's transfer function, T is the temperature, and T_0 is the temperature at calibration. The reference sensor whose reference input is s_0 generates voltage:

$$V_0 \approx [1 + \alpha(T - T_0)] f(s_0). \tag{5.81}$$

We consider ambient temperature as a transmitted multiplicative noise which affects both sensors in the same way. Taking a ratio of the above equations, we arrive at

$$\frac{V_1}{V_0} = \frac{1}{f(s_o)} f(s_1). \tag{5.82}$$

Since $f(s_0)$ is constant, the ratio is not temperature dependent. It should be emphasized, however, that the ratiometric technique is useful only when the anticipated noise has a multiplicative nature, whereas a differential technique works only for additive transmitted noise. Neither technique is useful for inherent noise, which is generated internally in sensors and circuits.

Although inherent noise is mostly Gaussian, the transmitted noise is usually less suitable for conventional statistical description. Transmitted noise may be periodic, irregularly recurring, or essentially random, and it ordinarily may be reduced substantially by taking precautions to minimize electrostatic and electromagnetic pickup from power sources at line frequencies and their harmonics, radio broadcast stations, arcing of mechanical switches, and current and voltage spikes resulting from switching in reactive (having inductance and capacitance) circuits. Such precautions may include filtering, decoupling, shielding of leads and components, use of guarding potentials, elimination of ground loops, physical reorientation of leads, components, and wires, use of damping diodes across relay coils and electric motors, choice of low impedances where possible, and choice of power supply and references having low noise. Transmitted noise from vibration may be reduced by proper mechanical design. A list outlining some of the sources of transmitted noise, their typical magnitudes, and some ways of dealing with them is shown in Table 5.4.

The most frequent channel for the coupling of electrical noise is a "parasitic" capacitance. Such a coupling exists everywhere. Any object is capacitively coupled to another object. For instance, a human standing on isolated earth develops a capacitance to ground on the order of 700 pF, electrical connectors have a pin-to-pin capacitance of about 2pF, and an optoisolator has an emitter-detector capacitance of about 2 pF. Figure 5.48A shows that an electrical noise source is connected to the sensor's internal impedance Z through a coupling capacitance C_S. That impedance

Table 5.4. Typical Sources of Transmitted Noise

External Source	Typical Magnitude	Typical Cure
60/50 Hz power	100 pA	Shielding; attention to ground loops; isolated power supply
120/100 Hz supply ripple	3 μV	Supply filtering
180/150 Hz magnetic pickup from saturated 60/50-Hz transformers	0.5 μV	Reorientation of components
Radio broadcast stations	1 mV	Shielding
Switch arcing	1 mV	Filtering of 5–100-MHz components; attention to ground loops and shielding
Vibration	10 pA (10–100 Hz)	Proper attention to mechanical coupling; elimination of leads with large voltages near input terminals and sensors
Cable vibration	100 pA	Use a low-noise (carbon-coated dielectric) cable
Circuit boards	0.01-10 pA/Hz below 10 Hz	Clean board thoroughly; use Teflon insulation where needed and guard well

Source: Adapted from Ref. [13].

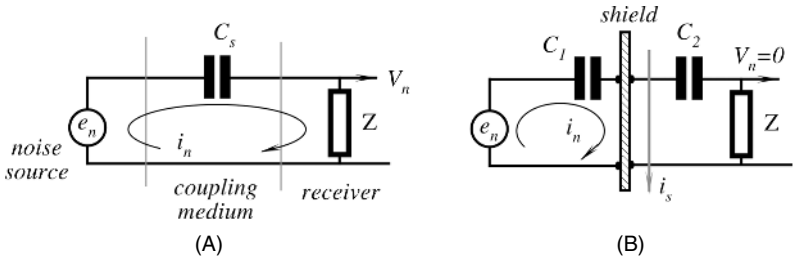


Fig. 5.48. Capacitive coupling (A) and electric shield (B).

may be a simple resistance or a combination of resistors, capacitors, inductors, and nonlinear elements, like diodes. Voltage across the impedance is a direct result of the change rate in the noise signal, the value of coupling capacitance C_s, and impedance Z. For instance, a pyroelectric detector may have an internal impedance which is equivalent to a parallel connection of a 30-pF capacitor and a 50-GΩ resistor. The sensor may be coupled through just 1 pF to a moving person who has the surface electrostatic charge on the body resulting in static voltage of 1000 V. If we assume that the main frequency of human movement is 1 Hz, the sensor would pick up an electrostatic interference of about 30 V! This is three to five orders of magnitude higher than the sensor would normally produce in response to thermal radiation received from the human body. Because some sensors and virtually all electronic circuits have nonlinearities, high-frequency interference signals, generally called RFI (radio-frequency interference) or EMI (electromagnetic interferences), may be rectified and appear at the output as a dc or slow-changing voltage.

5.9.3 Electric Shielding

Interferences attributed to electric fields can be significantly reduced by appropriate shielding of the sensor and circuit, especially of high impedance and nonlinear components. Each shielding problem must be analyzed separately and carefully. It is very important to identify the noise source and how it is coupled to the circuit. Improper shielding and guarding may only make matters worse or create a new problem.

A shielding serves two purposes [14]. First, it confines noise to a small region. This will prevent noise from getting into nearby circuits. However, the problem with such shields is that the noise captured by the shield can still cause problems if the return path that the noise takes is not carefully planned and implemented by an understanding of the ground system and making the connections correctly.

Second, if noise is present in the circuit, shields can be placed around critical parts to prevent the noise from getting into sensitive portions of the detectors and circuits. These shields may consist of metal boxes around circuit regions or cables with shields around the center conductors.

As it was shown in Section 3.1 of Chapter 3, the noise that resulted from the electric fields can be well controlled by metal enclosures because the charge q cannot exist on the interior of a closed conductive surface. Coupling by a mutual, or stray,

capacitance can be modeled by circuit shown in Fig. 5.48. Here, e_n is a noise source. It may be some kind of a part or component whose electric potential varies. C_s is the stray capacitance (having impedance Z_s at a particular frequency) between the noise source and the circuit impedance Z, which acts as a receiver of the noise. The voltage V_n is a result of the capacitive coupling. A noise current is defined as

$$i_n = \frac{V_n}{Z + Z_s} \tag{5.83}$$

and actually produces noise voltage

$$V_n = \frac{e_n}{1 + Z_c/Z}. \tag{5.84}$$

For example, if $C_s = 2.5$ pF, $Z = 10$ kΩ (resistor), and $e_n = 100$ mV, at 1.3MHz the output noise will be 20 mV.

One might think that 1.3-MHz noise is relatively easy to filter out from low-frequency signals produced by a sensor. In reality, it cannot be done, because many sensors and, especially the front stages of the amplifiers, contain nonlinear components (p-n-semiconductor junctions) which act as rectifiers. As a result, the spectrum of high-frequency noise shifts into a low-frequency region, making the noise signal similar to the voltage produced by a sensor.

When a shield is added, the change to the situation is shown in Fig. 5.48B. With the assumption that the shield has zero impedance, the noise current at the left side will be $i_{n_1} = e_n/Z_{C_1}$. On the other side of the shield, noise current will be essentially zero because there is no driving source on the right side of the circuit. Subsequently, the noise voltage over the receiving impedance will also be zero and the sensitive circuit becomes effectively shielded from the noise source. One must be careful, however, that there is no currents is flowing over the shield. Coupled with the shield resistance, these may generate additional noise. There are several practical rules that must be observed when applying electrostatic shields:

- An electrostatic shield, to be effective, should be connected to the reference potential of any circuitry contained within the shield. If the signal is connected to a ground (chassis of the frame or to earth), the shield must be connected to that ground. Grounding of shield is *useless* if the signal is not returned to the ground.
- If a shielding cable is used, its shield must be connected to the signal referenced node at the signal source side (Fig. 5.49A).
- If the shield is split into sections, as might occur if connectors are used, the shield for each segment must be tied to those for the adjoining segments and ultimately connected only to the signal referenced node (Fig. 5.49B).
- The number of separate shields required in a data acquisition system is equal to the number of independent signals that are being measured. Each signal should have its own shield, with no connection to other shields in the system, unless they share a common reference potential (signal "ground"). In that case, all connections must be made by a separate jumping wire connected to each shield at a single point.
- A shield must be grounded only at one point—preferably next to the sensor. A shielded cable must never be grounded at both ends (Fig. 5.50). The potential

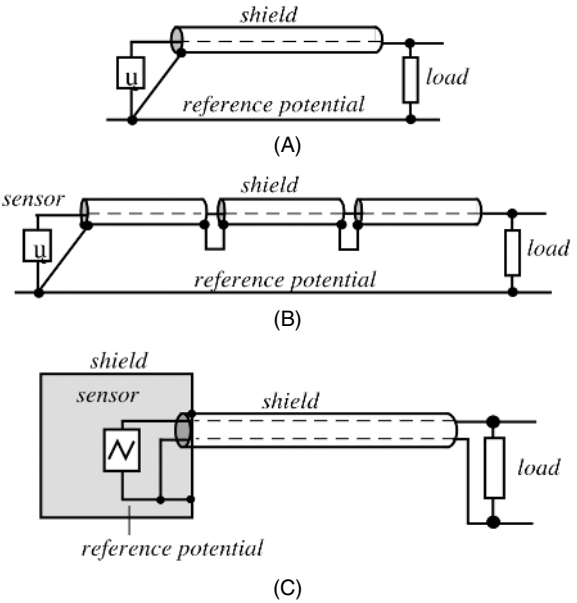


Fig. 5.49. Connections of an input cable to a reference potential.

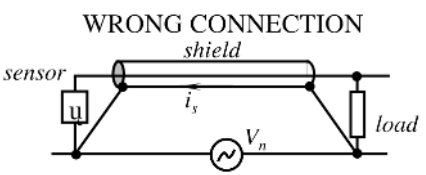


Fig. 5.50. Cable shield is erroneously grounded at both ends.

difference (V_n) between two "grounds" will cause the shield current i_s to flow which may induce a noise voltage into the center conductor via magnetic coupling.

- If a sensor is enclosed into a shield box and data are transmitted via a shielded cable (Fig. 5.49C), the cable shield must be connected to the box. It is a good practice to use a separate conductor for the reference potential ("ground") inside the shield, and not use the shield for any other purposes except shielding. Do not allow shield current to exist.
- Never allow the shield to be at any potential with respect to the reference potential (except in the case of driven shields, as shown in Fig. 5.4B). The shield voltage couples to the center conductor (or conductors) via a cable capacitance.
- Connect shields to a ground via short wires to minimize inductance. This is especially important when both analog and digital signals are transmitted.

5.9.4 Bypass Capacitors

The bypass capacitors are used to maintain a low power-supply impedance at the point of a load. Parasitic resistance and inductance in supply lines mean that the power-

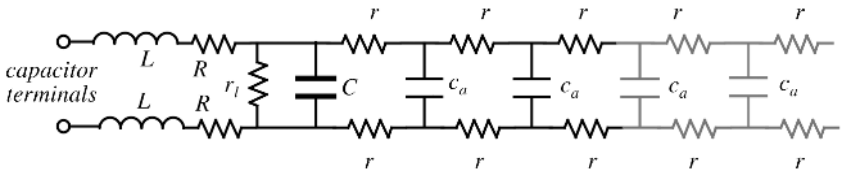


Fig. 5.51. Equivalent circuit of a capacitor.

supply impedance can be quite high. As the frequency increases, the inductive parasitic becomes troublesome and may result in circuit oscillation or ringing effects. Even if the circuit operates at lower frequencies, the bypass capacitors are still important, as high-frequency noise may be transmitted to the circuit and power-supply conductors from external sources, (e.g., radio stations). At high frequencies, no power supply or regulator has zero output impedance. What type of capacitor to use is determined by the application, frequency range of the circuit, cost, board space, and some other considerations. To select a bypass capacitor, one must remember that a practical capacitor at high frequencies may be far from the idealized capacitor described in textbooks.'

A generalized equivalent circuit of a capacitor is shown in Fig. 5.51. It is composed of a nominal capacitance C, leakage resistance r_l, lead inductances L, and resistances R. Further, it includes dielectric absorption terms r and c_a, which are manifested in the capacitor's "memory". In many interface circuits, especially amplifiers, analog integrators, and current (charge)-to-voltage converters, dielectric absorption is a major cause for errors. In such circuits, film capacitors should be used whenever possible.

In bypass applications, r_l and dielectric absorption are second-order terms, but series R and L are of importance. They limit the capacitor's ability to damp transients and maintain a low-power supply output impedance. Often, bypass capacitors must be of large values (10 μF or more) so they can absorb longer transients; thus, electrolytic capacitors are often employed. Unfortunately, these capacitors have large series R and L. Usually, tantalum capacitors offer better results; however, a combination of aluminum electrolytic with nonpolarized (ceramic or film) capacitors may offer even further improvement. A combination of the wrong types of bypass capacitor may lead to ringing, oscillation, and cross-talk between data communication channels. The best way to specify a correct combination of bypass capacitors is to first try them on a breadboard.

5.9.5 Magnetic Shielding

Proper shielding may dramatically reduce noise resulting from electrostatic and electrical fields. Unfortunately, it is much more difficult to shield against magnetic fields because it penetrates conducting materials. A typical shield placed around a conductor and grounded at one end has little, if any, effect on the magnetically induced voltage in that conductor. As a magnetic field B_0 penetrates the shield, its amplitude drops exponentially (Fig. 5.52B). The skin depth δ of the shield is the depth required for the field attenuation by 37% of that in the air. Table 5.5 lists typical values of δ for

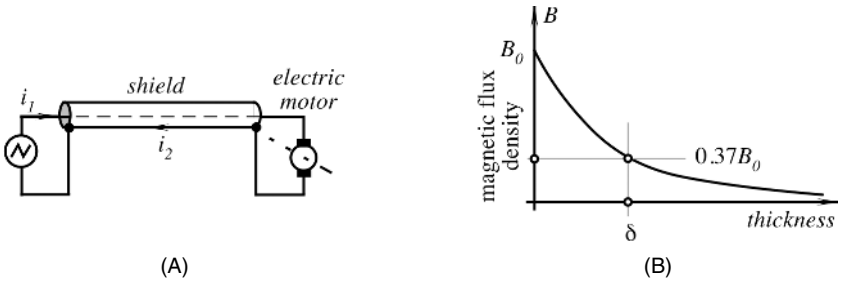


Fig. 5.52. Reduction of a transmitted magnetic noise by powering a load device through a coaxial cable (A); Magnetic shielding improves with the thickness of the shield (B).

Table 5.5. Skin Depth, δ, (in mm) Versus Frequency

Frequency	Copper	Aluminum	Steel
60 Hz	8.5	10.9	0.86
100 Hz	6.6	8.5	0.66
1 kHz	2.1	2.7	0.20
10 kHz	0.66	0.84	0.08
100 kHz	0.2	0.3	0.02
1 MHz	0.08	0.08	0.008

Source: Adapted from Ref. [15].

several materials at different frequencies. At high frequencies, any material may be used for effective shielding; however, at a lower range, steel yields a much better performance.

For improving low-frequency magnetic field shielding, a shield consisting of a high-permeability magnetic material (e.g., mumetal) should be considered. However, mumetal effectiveness drops at higher frequencies and strong magnetic fields. An effective magnetic shielding can be accomplished with thick steel shields at higher frequencies. Because magnetic shielding is very difficult, the most effective approach at low frequencies is to minimize the strength of magnetic fields, minimize the magnetic loop area at the receiving end, and select the optimal geometry of conductors. Some useful practical guidelines are as follows:

- Locate the receiving circuit as far as possible from the source of the magnetic field.
- Avoid running wires parallel to the magnetic field; instead, cross the magnetic field at right angles.
- Shield the magnetic field with an appropriate material for the frequency and strength.
- Use a twisted pair of wires for conductors carrying the high-level current that is the source of the magnetic field. If the currents in the two wires are equal and opposite, the net field in any direction over each cycle of twist will be zero. For this

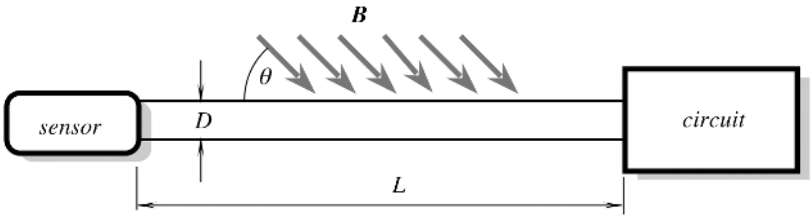


Fig. 5.53. Receiver's loop is formed by long conductors.

arrangement to work, none of the current can be shared with another conductor, (e.g., a ground plane, which may result in ground loops).

- Use a shielded cable with the high-level source circuit's return current carried by the shield (Fig. 5.52A). If the shield current i_2 is equal and opposite to that of the center conductor i_1, the center conductor field and the shield field will cancel, producing a zero net field. This case seems to be a violation of the rule "no shield currents" for the receiver's circuit; however, the shielded cable here is not used to electrostatically shield the center conductor. Instead, the geometry produces a cancellation of the magnetic field which is generated by a current supplied to a "current-hungry" device (an electric motor in this example)
- Because magnetically induced noise depends on the area of the receiver loop, the induced voltage due to magnetic coupling can be reduced by making the loop's area smaller.

What is the receiver's loop? Figure 5.53 shows a sensor which is connected to the load circuit via two conductors having length L and separated by distance D. The rectangular circuit forms a loop area $a = LD$. The voltage induced in series with the loop is proportional to the area and the cosine of its angle to the field. Thus, to minimize noise, the loop should be oriented at right angles to the field and its area should be minimized.

The area can be decreased by reducing the length of the conductors and/or decreasing the distance between the conductors. This is easily accomplished with a twisted pair, or at least with a tightly cabled pair of conductors. It is good practice to pair the conductors so that the circuit wire and its return path will always be together. This requirement must not be overlooked. For instance, if wires are correctly positioned by a designer, a service technician may reposition them during the repair work. A new wire location may create a disastrous noise level. Hence, a general rule is: Know the area and orientation of the wires and permanently secure the wiring.

Magnetic fields are much more difficult to shield against than electric fields because they can penetrate conductive materials.

5.9.6 Mechanical Noise

Vibration and *acceleration effects* are also sources of transmitted noise in sensors which otherwise should be immune to them. These effects may alter transfer charac-

teristics (multiplicative noise) or they may result in the generation of spurious signals (additive noise) by a sensor. If a sensor incorporates certain mechanical elements, vibration along some axes with a given frequency and amplitude may cause resonant effects. For some sensors, acceleration is a source of noise. For instance, most pyroelectric detectors also possess piezoelectric properties. The main function of the detector is to respond to thermal gradients. However, such environmental mechanical factors as fast changing air pressure, strong wind, or structural vibration cause the sensor to respond with output signals which often are indistinguishable from responses to normal stimuli.

5.9.7 Ground Planes

For many years, ground planes have been known to electronic engineers and printed circuit designers as a "mystical and ill-defined" cure for spurious circuit operation [16]. Ground planes are primarily useful for minimizing circuit inductance. They do this by utilizing the basic magnetic theory. Current flowing in a wire produces an associated magnetic field (Section 3.3 of Chapter 3). The field's strength is proportional to the current i and inversely related to the distance r from the conductor:

$$B = \frac{\mu_0 i}{2\pi r}. \tag{5.85}$$

Thus, we can imagine a current carrying wire surrounded by a magnetic field. Wire inductance is defined as energy stored in the field set up by the wire's current. To compute the wire's inductance requires integrating the field over the wire's length and the total area of the field. This implies integrating on the radius from the wire surface to infinity. However, if two wires carrying the same current in opposite directions are in close proximity, their magnetic fields are canceled. In this case, the virtual wire inductance is much smaller. An opposite flowing current is called *return current*. This is the underlying reason for ground planes. A ground plane provides a return path directly under the signal-carrying conductor through which return current can flow. Return current has a direct path to ground, regardless of the number of branches associated with the conductor. Currents will always flow through the return path of the lowest impedance. In a properly designed ground plane, this path is directly under the signal conductor. In practical circuits, a ground plane is on one side of the board and the signal conductors are on the other. In the multilayer boards, a ground plane is usually sandwiched between two or more conductor planes. Aside from minimizing parasitic inductance, ground planes have additional benefits. Their flat surface minimizes resistive losses due to the "skin effect" (ac current travel along a conductor's surface). Additionally, they aid the circuit's high-frequency stability by referring stray capacitance to the ground. Some practical suggestions are as follows:

- Make ground planes of as much area as possible on the component side (or inside for the multilayer boards). Maximize the area especially under traces that operate with high frequency or digital signals.
- Mount components that conduct fast transient currents (terminal resistors, ICs, transistors, decoupling capacitors, etc.) as close to the board as possible.

- Wherever a common ground reference potential is required, use separate conductors for the reference potential and connect them all to the ground plane at a common point to avoid voltage drops due to ground currents.
- Keep the trace length short. Inductance varies directly with length and no ground plane will achieve perfect cancellation.

5.9.8 Ground Loops and Ground Isolation

When a circuit is used for low-level input signals, a circuit itself may generate enough noise to present a substantial problem for accuracy. Sometimes, a circuit is correctly designed on paper, and a bench breadboard shows a quite satisfactory performance; however, when a production prototype with the printed circuit board is tested, the accuracy requirement is not met. A difference between a breadboard and PC-board prototypes may be in the physical layout of conductors. Usually, conductors between electronic components are quite specific: They may connect a capacitor to a resistor, a gate of a JFET transistor to the output of an operational amplifier, and so forth. However, there are at least two conductors which, in most cases, are common for the majority of the electronic circuit. These are the power-supply bus and the ground bus. Both may carry undesirable signals from one part of the circuit to another; specifically, they may couple strong output signals to the sensitive input stages.

A power-supply bus carries supply currents to all stages. A ground bus also carries supply currents, but, in addition, it is often used to establish a reference base for an electrical signal. Interaction of these two functions may lead to a problem which is known as a ground loop. We illustrate it in Fig. 5.54A in which a sensor is connected to a positive input of an amplifier which may have a substantial gain. The amplifier is connected to the power supply and draws current i which is returned to the ground bus as i'. A sensor generates voltage V_s which is fed to the positive input of the amplifier. A ground wire is connected to the circuit in point a—right next to the sensor's terminal. A circuit has no visible error sources, nevertheless, the output voltage contain substantial errors. A noise source is developed in a wrong connection of ground wires. Figure 5.54B shows that the ground conductor is not ideal. It may have some finite resistance R_g and inductance L_g. In this example, supply current while returning to the battery from the amplifier passes through the ground bus between points b and a resulting in voltage drop V_g. This drop, however small, may be comparable with the signal produced by the sensor. It should be noted that voltage V_g is serially connected with the sensor and is directly applied to the amplifier's input. Ground currents may also contain high-frequency components; then, the bus inductance will produce quite strong spurious high-frequency signals which not only add noise to the sensor but may cause circuit instability as well. For example, let us consider a thermopile sensor which produces voltage corresponding to 100 μV/°C of the object's temperature. A low-noise amplifier has quiescent current $i = 5$ mA, which passes through the ground loop having resistance $R_g = 0.2\Omega$. The ground-loop voltage $V_g = i R_g = 1$ mV corresponds to an error of -10°C! The cure is usually quite simple: Ground loops must be broken. A circuit designer should always separate a reference ground from current-carrying grounds, especially serving digital devices. Figure 5.55 shows

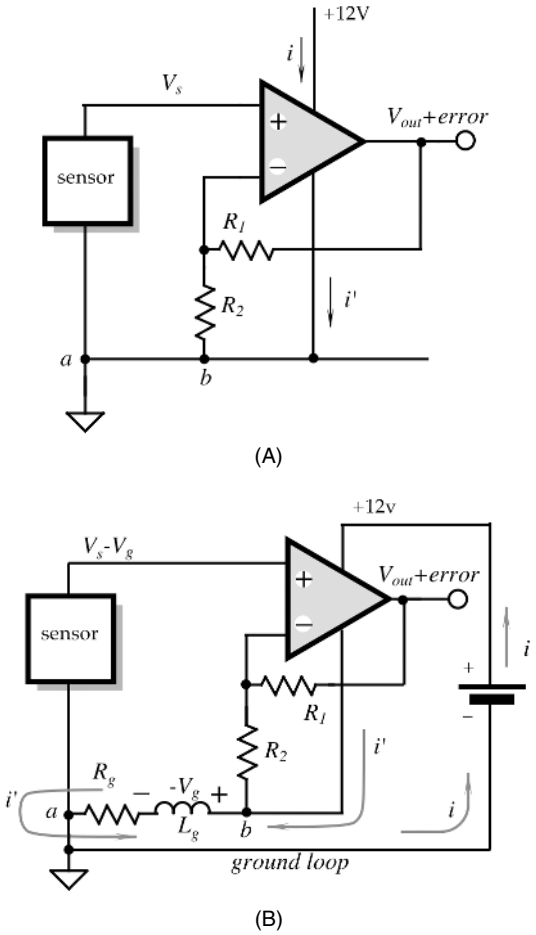


Fig. 5.54. Wrong connection of a ground terminal to a circuit (A); path of a supply current through the ground conductors (B).

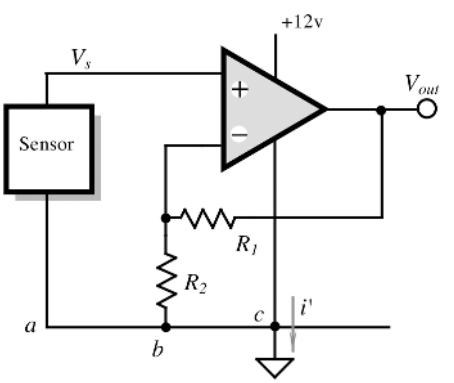


Fig. 5.55. Correct grounding of a sensor and interface circuit.

that moving the ground connection from sensor's point *a* to the power terminal point *c* prevents formation of spurious voltage across the ground conductor connected to the sensor and a feedback resistor R_2. A rule of thumb is to connect the ground to the circuit board at only one point. Grounding at two or more spots may form ground loops; which often is very difficult to diagnose.

5.9.9 Seebeck Noise

This noise is a result of the Seebeck effect (Section 3.9 of Chapter 3) which is manifested as the generation of an electromotive force (e.m.f.) when two dissimilar metals are joined together. The Seebeck e.m.f. is small and, for many sensors, may be simply ignored. However, when absolute accuracy on the order of 10–100 μV is required, that noise must be taken into account. The connection of two dissimilar metals produces a temperature sensor. However, when temperature sensing is not a desired function, a thermally induced e.m.f. is a spurious signal. In electronic circuits, the connection of dissimilar metals can be found everywhere: connectors, switches, relay contacts, sockets, wires, and so on. For instance, the copper PC board cladding connected to Kovar[TM6] input pins of an integrated circuit creates an offset voltage of 40 $\mu V \cdot \Delta T$ where ΔT is the temperature gradient (in °C) between two dissimilar metal contacts. The common lead–tin solder, when used with the copper cladding creates a thermoelectric voltage between 1 and 3 μV/°C. There are special cadmium–tin solders available to reduce these spurious signals down to 0.3 μV/°C. Figure 5.56A shows the Seebeck e.m.f. for two types of solder. The connection of two identical wires fabricated by different manufacturers may result in the voltage having a slope on the order of 200 nV/°C.

In many cases, Seebeck e.m.f. may be eliminated by a proper circuit layout and thermal balancing. It is a good practice to limit the number of junctions between the sensor and the front stage of the interface circuit. Avoid connectors, sockets, switches, and other potential sources of e.m.f. to the extent possible. In some cases, this will not be possible. In these instances, attempt to balance the number and type of junctions in the circuit's front stage so that differential cancellations occur. Doing this may involve deliberately creating and introducing junctions to offset necessary junctions. Junctions which the intent to produce cancellations must be maintained at the same temperature. Figure 5.56B shows a remote sensor connection to an amplifier where the sensor junctions, input terminal junctions, and amplifier components junctions are all maintained at different but properly arranged temperatures. Such thermally balanced junctions must be maintained at close physical proximity and preferably on common heat sinks. Air drafts and temperature gradients in the circuit boards and sensor enclosures must be avoided.

[6] Trademark of Westinghouse Electric Corp.

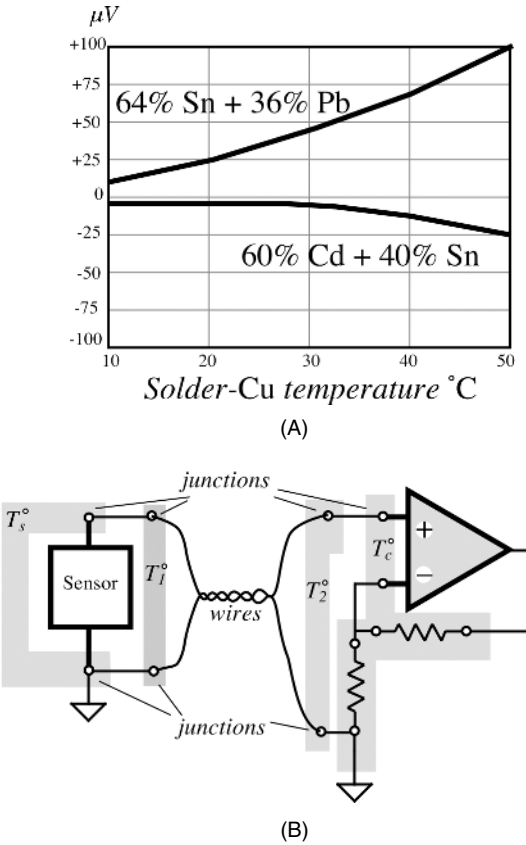


Fig. 5.56. (A) Seebeck e.m.f. developed by solder–copper joints (Adapted from Ref. [17]); (B) Maintaining joints at the same temperature reduces Seebeck noise.

5.10 Batteries for Low Power Sensors

Modern development of integrated sensors and need for long-term remote monitoring and data acquisition demand the use of reliable and high-energy density power sources. The history of battery development goes back to Volta and shows a remarkable progress during the last decades. Well-known old electrochemical power sources improved dramatically. Examples are C–Zn, alkaline, Zn–air, NiCd, and lead–acid batteries. Currently, newer systems such as secondary Zn–air, Ni–metal hydride, and, especially, lithium batteries are growing in use as new devices are designed around their higher voltage and superior shelf life. The Li–MnO_2 system dominates the commercial market where they range from miniature flat cells to "D" size cells.

All batteries can be divided into two groups: *primary* (single use devices) and *secondary* (rechargeable) (multiple-use devices).

Often, batteries are characterized by energy per unit weight, however, for miniature sensor applications energy per unit volume often becomes more critical.

Table A.20 (Appendix) shows typical characteristics of the carbon–zinc and alkaline cells (power density in watt·hour per liter and per kilogram.)

In general, the energy delivered by a battery depends on the rate at which power is withdrawn. Typically, as the current is increased, the amount of energy delivered is decreased. Battery energy and power are also affected by the construction of the battery, the size, and the duty cycle of current delivery. The manufacturers usually specify batteries as ampere-hours or watt-hours when discharged at a specific rate to a specific voltage cutoff. For instance, if the battery capacitance is C (in mA·h) and the average current drain is I (mA), the time of a battery discharge (lifetime for a primary cell) is defined as

$$t = \frac{C}{In}, \tag{5.86}$$

where n is the duty cycle. For instance, if the battery is rated as having capacity of 50 mA h, the circuit operating current consumption is about 5 mA, and the circuit works only 5 min every hour (duty cycle is 5/60), the battery will last for

$$t = \frac{C}{In} = \frac{(50)(60)}{(5)(5)} = 120 \text{ h}$$

Yet, the manufacturer's specification must be used with a grain of salt and *only as a guideline*, for the specified discharge rate rarely coincides with the actual power consumption. It is highly recommended to determine the battery life experimentally, rather than rely on the calculation. When designing the electronic circuit, its power consumption shall be determined during various operating modes and over the operating temperature range. Then, these values of power consumption must be used in the simulation of the battery load to determine the useful life with a circuit-specific cutoff voltage in mind. Sometimes, a circuit draws high currents during short times (pulse mode) and the battery's ability to deliver such a pulse current should be evaluated. If a battery cannot deliver a high pulse current, a parallel electrolytic capacitor serving as a storage tank may be considered.

It should be noted that the accelerated life tests of a battery must be used with caution, because as it was noted earlier, the useful capacity of a battery greatly depends on the load, operational current profile, and duty cycle.

5.10.1 Primary Cells

The construction of a battery cell determines its performance and cost. Most primary cells employ single, thick electrodes arranged in a parallel or concentric configuration and aqueous electrolytes. Most small secondary cells are designed differently; they use a "wound" or "jelly roll" construction, in which long, thin electrodes are wound into a cylinder and placed into a metal container. This results in a higher power density, but decreased energy density and higher cost. Due to the low conductivity of electrolytes, many lithium primary cells also use the "wound" construction [18].

Leclanche (Carbon–Zinc) Batteries. These batteries use zinc as the anode. They are of two types. One uses natural manganese dioxide as the cathode with an ammonium chloride electrolyte. A "premium" version uses electrolytic manganese dioxide as the cathode and a zinc chloride electrolyte. These batteries are still the most popular worldwide, especially in the Orient, being produced by over 200 manufacturers. Their use is about equal to that of the alkaline in Europe, but is only near 25% of alkaline batteries in the United States. These batteries are preferred when a high power density is not required and shelf life is not critical, but the low cost is a dominating factor.

Alkaline Manganese Batteries. Demand for these batteries grew significantly, especially after a major improvement: the elimination of mercury from the zinc anode. The alkaline batteries are capable of delivering high currents, have improved power/density ratio, and have at least 5 years of shelf life (Table A.20).

Primary Lithium batteries. Most of these batteries are being produced in Japan. The popularity of lithium–manganese dioxide cells grows rapidly thanks to their higher operating voltage, wide range of sizes and capacities, and excellent shelf life (Table A.21). Lithium iodine cells have a very high energy density and allow up to 10 years of operation in a pacemaker (implantable heart rate controller). However, these batteries are designed with a low-conductivity solid-state electrolyte and allow operation with very low current drain (on the order of microamperes), which often is quite sufficient in cases for which passive sensors are employed.

The amount of lithium in the batteries is quite small, because just 1 g is sufficient for producing a capacity of 3.86 A h. Lithium cells are exempt from environmental regulations, but are still considered hazardous because of their flammability.

5.10.2 Secondary Cells

Secondary cells (Tables A.22 and A.23) are rechargeable batteries. Sealed lead acid batteries offer small size at large capacities and allow about 200 cycles of life at discharge times as short as 1 h. The main advantages of these cells are low initial cost, low self-discharge, on the ability to support heavy loads and withstand harsh environments. In addition, these batteries have a long life. The disadvantages include relatively large size and weight as well as potential environmental hazard due to the presence of lead and sulfuric acid.

Sealed *nickel–cadmium* (NiCd) and *nickel–metal hydrate* (Ni-MH) are the most widely used secondary cells, being produced at volumes of over 1 billion cells per year. The typical capacity for an "AA" cell is about 800 mA h and even higher from some manufacturers. This is possible thanks to the use of a high-porosity nickel foam or felt instead of traditional sintered nickel as the carrier for the active materials. The NiCd cells are quite tolerant of overcharge and overdischarge. An interesting property of NiCd is that charging is an endothermic process, (i.e., the battery absorbs heat), whereas other batteries warm up when charging. Cadmium, however, presents a potential environmental problem. Bi-MH and modern NiCd do not exhibit a "memory" effect; that is, partial discharge does not influence their ability to fully recharge. The

nickel–metal hydrate cells is nearly direct replacement for NiCd, yet they yield better capacity but have somewhat poorer self-discharge.

A lithium polymer battery contain a nonliquid electrolyte, which makes it a solid-state battery. This allows one to fabricate it in any size and shape; however, these batteries are the most expensive. Rechargeable alkaline batteries have low cost and good power density. However, their life cycles are quite low.

References

1. Widlar, R. J. Working with high impedance Op Amps. In: *Linear Application Handbook*. National Semiconductor, 1980.
2. Pease, R. A. Improve circuit performance with a 1-op-amp current pump. Electronic Design News, 85–90, Jan. 20, 1983.
3. Bell, D.A. *Solid State Pulse Circuits*, 2nd ed. Reston, Reston, VA, 1981.
4. Sheingold, D. H., ed. *Analog–Digital Conversion Handbook*, 3rd ed. Prentice-Hall, Englewood Cliffs, NJ, 1986.
5. Williams, J. Some techniques for direct digitization of transducer outputs, In: *Linear Technology Application Handbook*, National Semiconductor, 1990.
6. Park, Y. E. and Wise, K. D. An MOS switched-capacitor readout amplifier for capacitive pressure sensors. *IEEE Custom IC Conference*, 1983, pp. 380–384.
7. Stafford, K.R., Gray, P. R., and Blanchard, R.A. A complete monolithic sample/hold amplifier. IEEE *J. of Solid-State Circuits*, 9, pp: 381–387, Dec. 1974.
8. Cho S. T. and Wise, K. D. A self-testing ultrasensitive silicon microflow sensor. *Sensor Expo Proceedings*, 1991, p. 208B-1.
9. Weatherwax, S. Understanding constant voltage and constant current excitation for pressure sensors. *SenSym Solid-State Sensor Handbook*. © Sensym, Inc., 1991.
10. Coats, M. R. New technology two-wire transmitters. *Sensors* 8(1), 1991.
11. Sheingold, D. H., ed. *Nonlinear Circuits Handbook*. Analog Devices, Inc. Northwood, MA, 1974.
12. Johnson, J. B. Thermal agitation of electricity in conductors. *Phys. Rev.* 1928.
13. *The Best of Analog Dialogue*. Analog Devices, Inc., Northwood, MA, 1991.
14. Rich, A. Shielding and guarding. In: *The Best of Analog Dialogue*. Analog Devices, Inc., Northwood, MA, 1991.
15. Ott, H. W. *Noise Reduction Techniques in Electronic Systems*. John Wiley & Sons, New York, 1976.
16. Williams, J. High speed comparator techniques. In: *Linear Applications Handbook*. Linear Technology Corp., 1990.
17. Pascoe, G. The choice of solders for high-gain devices. *New Electron. (U.K.)* 1977.
18. Powers R.A. Batteries for low power electronics. In: *Proc. IEEE* 83(4), 687–693, 1995.

This page intentionally left blank

6

Occupancy and Motion Detectors

September 11, 2001 has changed the way people think about airport, aviation, and security in general. The threat is expanding interest in more reliable systems to detect the presence of people within the protected perimeters. The *occupancy* sensors detect the presence of people (and sometimes animals) in a monitored area. *Motion detectors* respond only to moving objects. A distinction between the two is that the occupancy sensors produce signals whenever an object is stationary or not, whereas the motion detectors are selectively sensitive to moving objects. The applications of these sensors include security, surveillance, energy management, (electric lights control), personal safety, friendly home appliances, interactive toys, novelty products, and so forth. Depending on the applications, the presence of humans may be detected through any means associated with some kind of a human body's property or body's actions [1]. For instance, a detector may be sensitive to body weight, heat, sounds, dielectric constant, and so forth. The following types of detector are presently used for the occupancy and motion sensing of people:

1. *Air pressure sensors*: detects changes in air pressure resulted from opening doors and windows
2. *Capacitive*: detectors of human body capacitance
3. *Acoustic*: detectors of sound produced by people
4. *Photoelectric*: interruption of light beams by moving objects
5. *Optoelectric*: detection of variations in illumination or optical contrast in the protected area
6. *Pressure mat switches*: pressure-sensitive long strips used on floors beneath the carpets to detect weight of an intruder
7. *Stress detectors*: strain gauges imbedded into floor beams, staircases, and other structural components
8. *Switch sensors*: electrical contacts connected to doors and windows
9. *Magnetic switches*: a noncontact version of switch sensors
10. *Vibration detectors*: react to the vibration of walls or other building structures, also may be attached to doors or windows to detect movements
11. *Glass breakage detectors*: sensors reacting to specific vibrations produced by shattered glass

12. *Infrared motion detectors*: devices sensitive to heat waves emanated from warm or cold moving objects
13. *Microwave detectors*: active sensors responsive to microwave electromagnetic signals reflected from objects
14. *Ultrasonic detectors*: similar to microwaves except that instead of electromagnetic radiation, ultrasonic waves are used
15. *Video motion detectors*: video equipment which compares a stationary image stored in memory with the current image from the protected area
16. *Video face recognition system*: image analyzers that compare facial features with a database
17. *Laser system detectors*: similar to photoelectric detectors, except that they use narrow light beams and combinations of reflectors
18. *Triboelectric detectors*: sensors capable of detecting static electric charges carried by moving objects

One of the major aggravations in detecting the occupancy or intrusion is a *false-positive* detection. The term "false positive" means that the system indicates an intrusion when there is none. In some noncritical applications where false-positive detections occur occasionally, (e.g., in a toy or a motion switch controlling electric lights in a room), this may be not a serious problem: The lights will be erroneously turned on for a short time, which unlikely do any harm.[1] In other systems, especially used for security and military purposes, the false-positive detections, although generally not as dangerous as false-negative ones (missing an intrusion), may become a serious problem. While selecting a sensor for critical applications, considerations should be given to its reliability, selectivity, and noise immunity. It is often a good practice to form a multiple-sensor arrangement with symmetrical interface circuits. It may dramatically improve the reliability of a system, especially in the presence of external transmitted noise. Another efficient way to reduce erroneous detections is to use sensors operating on different physical principles [2] (e.g., combining capacitive and infrared detectors is an efficient combination, as they are receptive to different kinds of transmitted noise.

6.1 Ultrasonic Sensors

These detectors are based on the transmission to the object and receiving reflected acoustic waves. A description of the ultrasonic detectors can be found in Section 7.6 of Chapter 7. For the motion detectors, they may require a somewhat longer operating range and a wider angle of coverage.

6.2 Microwave Motion Detectors

The microwave detectors offer an attractive alternative to other detectors when it is required to cover large areas and to operate over an extended temperature range under the influence of strong interferences, such as wind, acoustic noise, fog, dust,

[1] Perhaps just stirring up some agitation about the presence of a ghost.

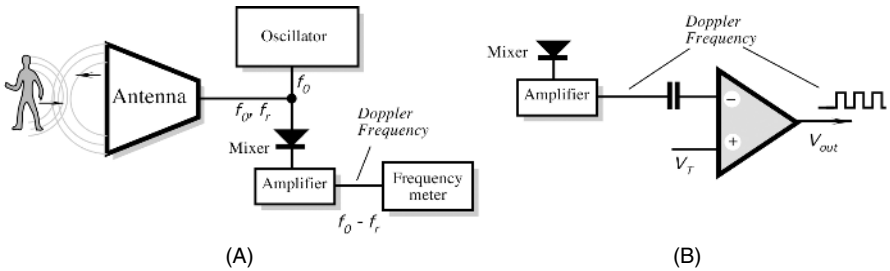


Fig. 6.1. Microwave occupancy detector: (A) a circuit for measuring Doppler frequency; (B) a circuit with a threshold detector.

moisture, and so forth. The operating principle of the microwave detector is based on radiation of electromagnetic radio-frequency (RF) waves toward a protected area. The most common frequencies are 10.525 GHz (X band) and 24.125 GHz (K band).[2] These wavelengths are long enough ($\lambda = 3$ cm at X band) to pass freely through most contaminants, such as airborne dust, and short enough for being reflected by larger objects.

The microwave part of the detector consists of a Gunn oscillator, an antenna, and a mixer diode. The Gunn oscillator is a diode mounted in a small precision cavity which, upon application of power, oscillates at microwave frequencies. The oscillator produces electromagnetic waves (frequency f_0), part of which is directed through an iris into a waveguide and focusing antenna which directs the radiation toward the object. Focusing characteristics of the antenna are determined by the application. As a general rule, the narrower the directional diagram of the antenna, the more sensitive it is (the antenna has a higher gain). Another general rule is that a narrow-beam antenna is much larger, whereas a wide-angle antenna can be quite small. The typical radiated power of the transmitter is 10–20 mW. A Gunn oscillator is sensitive to the stability of applied dc voltage and, therefore, must be powered by a good quality voltage regulator. The oscillator may run continuously, or it can be pulsed, which reduces the power consumption from the power supply.

The smaller part of the microwave oscillations is coupled to the Schottky mixing diode and serves as a reference signal (Fig. 6.1A). In many cases, the transmitter and the receiver are contained in one module called a transceiver. The target reflects some waves back toward the antenna, which directs the received radiation toward the mixing diode whose current contains a harmonic with a phase differential between the transmitted and reflected waves. The phase difference is in a direct relationship to the distance to the target. The *phase-sensitive* detector is useful mostly for detecting the distance to an object. However, movement, not distance, should be detected. Thus, for the occupancy and motion detector, the Doppler effect is the basis for the operation of microwave and ultrasonic detectors. It should be noted that the Doppler-effect device is a true motion detector because it is responsive only to moving targets. Here is how it works.

[2] The power of radiation must be sufficiently low not to present any health hazards.

An antenna transmits the frequency f_0 which is defined by the wavelength λ_0 as

$$f_0 = \frac{c_0}{\lambda_0}, \tag{6.1}$$

where c_0 is the speed of light. When the target moves toward or away from the transmitting antenna, the frequency of the reflected radiation will change. Thus, if the target is moving away with velocity v, the reflected frequency will decrease and it will increase for the approaching targets. This is called the *Doppler effect*, after the Austrian scientist Christian Johann Doppler (1803–1853).[3] Although the effect first was discovered for sound, it is applicable to electromagnetic radiation as well. However, in contrast to sound waves that may propagate with velocities dependent on the movement of the source of the sound, electromagnetic waves propagate with the speed of light, which is an absolute constant. The frequency of reflected electromagnetic waves can be predicted by the theory of relativity as

$$f_r = f_0 \frac{\sqrt{1-(v/c_0)^2}}{1+v/c_0}. \tag{6.2}$$

For practical purposes, however, the quantity $(v/c_0)^2$ is very small as compared with unity; hence, it can be ignored. Then, the equation for the frequency of the reflected waves becomes identical to that for the acoustic waves:

$$f_r = f_0 \frac{1}{1+v/c_0}. \tag{6.3}$$

Due to a Doppler effect, the reflected waves have a different frequency f_r. The mixing diode combines the radiated (reference) and reflected frequencies and, being a nonlinear device, produces a signal which contains multiple harmonics of both frequencies. The electric current through the diode may be represented by a polynomial:

$$i = i_0 + \sum_{k=1}^{n} a_k (U_1 \cos 2\pi f_0 t + U_2 \cos 2\pi f_r t)^k, \tag{6.4}$$

where i_0 is a dc component, a_k are harmonic coefficients which depend on a diode operating point, U_1 and U_2 are amplitudes of the reference and received signals, respectively, and t is time. The current through a diode contains an infinite number of harmonics, among which there is an harmonic of a differential frequency: $a_2 U_1 U_2 \cos 2\pi (f_0 - f_r)t$, which is called a Doppler frequency Δf.

The Doppler frequency in the mixer can be found from Eq. (6.3):

$$\Delta f = f_0 - f_r = f_0 \frac{1}{c_0/v + 1}; \tag{6.5}$$

[3] One hundred fifty years ago acoustical instruments for precision measurements were not available; yet, to prove his theory, Doppler placed trumpeters on a railroad flatcar and musicians with a sense of absolute pitch near the tracks. A locomotive engine pulled the flatcar back and forth at different speeds for two days. The musicians on the ground "recorded" the trumpet notes as the train approached and receded. The equations held up.

and since $c_0/v \gg 1$, the following holds after substituting Eq. (6.1):

$$\Delta f \approx \frac{v}{\lambda_0}. \tag{6.6}$$

Therefore, the signal frequency at the output of the mixer is linearly proportional to the velocity of a moving target. For instance, if a person walks toward the detectors with a velocity of 0.6 m/s, a Doppler frequency for the X-band detector is $\Delta f = 0.6/0.03 = 20$ Hz.

Equation (6.6) holds true only for movements in the normal direction. When the target moves at angles Θ with respect to the detector, the Doppler frequency is

$$\Delta f \approx \frac{v}{\lambda_0} \cos \Theta. \tag{6.7}$$

This implies that Doppler detectors theoretically become insensitive when a target moves at angles approaching 90°. In the velocity meters, to determine the velocity of a target, it is required is to measure the Doppler frequency and the phase to determine the direction of the movement (Fig. 6.1A). This method is used in police radars. For supermarket door openers and security alarms, instead of measuring the frequency, a threshold comparator is used to indicate the presence of a moving target (Fig. 6.1B). It should be noted that even if Eq. (6.7) predicts that the Doppler frequency is near zero for targets moving at angles $\Theta = 90°$, the entering of a target into the protected area at any angle results in an abrupt change in the received signal amplitude, and the output voltage from the mixer changes accordingly. Usually, this is sufficient to trigger the response of a threshold detector.

The signal from the mixer is in the range from microvolts to millivolts, so the amplification is needed for signal processing. Because the Doppler frequency is in the audio range, the amplifier is relatively simple; however, it generally must be accompanied by so-called notch filters, which reject a power line frequency and the main harmonic from full-wave rectifiers and fluorescent light fixtures: 60 and 120 Hz (or 50 and 100 Hz). For the normal operation, the received power must be sufficiently high. It depends on several factors, including the antenna aperture area A, target area a, and distance to the target r:

$$P_r = \rho \frac{P_0 A^2 a}{4\pi \lambda^2 r^4}, \tag{6.8}$$

where P_0 is the transmitted power. For effective operation, the target's cross-sectional area a must be relatively large, because for $\lambda^2 \leq a$, the received signal is drastically reduced. Further, the reflectivity ρ of a target in the operating wavelength is also very important for the magnitude of the received signal. Generally, conductive materials and objects with high dielectric constants are good reflectors of electromagnetic radiation, whereas many dielectrics absorb energy and reflect very little. Plastics and ceramics are quite transmissive and can be used as windows in the microwave detectors. The best target for a microwave detector is a smooth, flat conductive plate positioned normally toward the detector. A flat conductive surface makes a very good

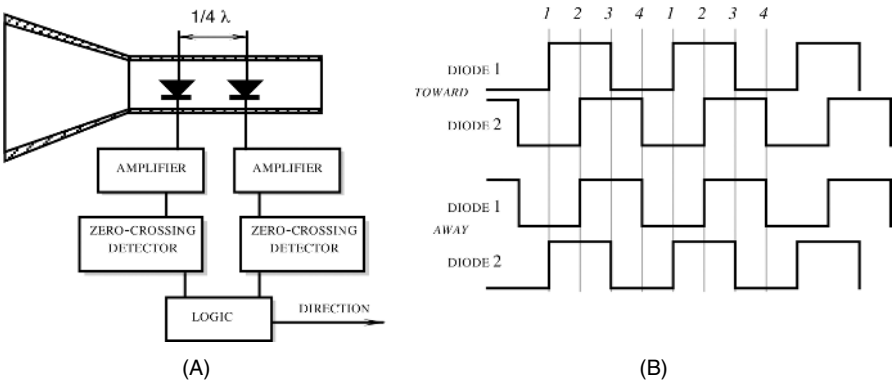


Fig. 6.2. Block diagram (A) and timing diagrams (B) of a microwave Doppler motion detector with directional sensitivity.

reflector; however, it may render the detector inoperable at angles other than 0°. Thus, an angle of $\Theta = 45°$ can completely divert a reflective signal from the receiving antenna. This method of diversion was used quite effectively in the design of the Stealth bomber, which is virtually invisible on radar screens.

To detect whether a target moves toward or away from the antenna, the Doppler concept can be extended by adding another mixing diode to the transceiver module. The second diode is located in the waveguide in such a manner that the Doppler signals from both diodes differ in phase by one-quarter of wavelength or by 90° (Fig. 6.2A). These outputs are amplified separately and converted into square pulses which can be analyzed by a logic circuit. The circuit is a digital phase discriminator that determines the direction of motion (Fig. 6.2B). Door openers and traffic control are two major applications for this type of module. Both applications need the ability to acquire a great deal of information about the target for discrimination before enabling a response. In door openers, limiting the field of view and transmitted power may substantially reduce the number of false-positive detections. Although for door openers a direction discrimination is optional, for traffic control it is a necessity to reject signals from the vehicles moving away. If the module is used for intrusion detection, the vibration of building structures may cause a large number of false-positive detections. A direction discriminator will respond to vibration with an alternate signal, and it will respond to an intruder with a steady logic signal. Hence, the direction discriminator is an efficient way to improve the reliability of the detection.

Whenever a microwave detector is used in the United States, it must comply with the strict requirements (e.g., MSM20100) imposed by the Federal Communication Commission. Similar regulations are enforced in many other countries. Also, the emission of the transmitter must be below 10 mW/cm^2 as averaged over any 0.1-h period, as specified by OSHA 1910.97 for the frequency range from 100 MHz to 100 GHz.

A quite effective motion detector may be designed by employing micropower impulse radar (see Section 7.7.1 of Chapter 7). Advantages of such detectors are very low power consumption and nearly total invisibility to the intruder. The radar may

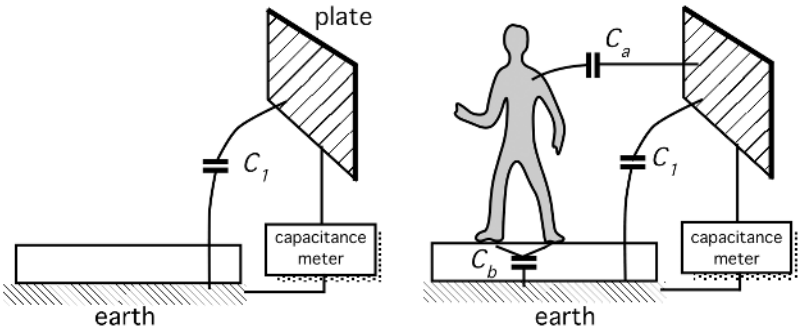


Fig. 6.3. An intruder brings in an additional capacitance to a detection circuit.

be concealed inside wooden or masonic structures and is virtually undetectable by electronic means thanks to its low emission resembling natural thermal noise.

6.3 Capacitive Occupancy Detectors

Being a conductive medium with a high dielectric constant, a human body develops a coupling capacitance to its surroundings.[4] This capacitance greatly depends on such factors as body size, clothing, materials, type of surrounding objects, weather, and so forth. However wide the coupling range is, the capacitance may vary from a few picofarads to several nanofarads. When a person moves, the coupling capacitance changes, thus making it possible to discriminate static objects from the moving ones. In effect, all objects form some degree of a capacitive coupling with respect to one another. If a human (or for that purpose, anything) moves into the vicinity of the objects whose coupling capacitance with each other has been previously established, a new capacitive value arises between the objects as a result of the presence of an intruding body. Figure 6.3 shows that the capacitance between a test plate and earth[5] is equal to C_1. When a person moves into the vicinity of the plate, it forms two additional capacitors: one between the plate and its own body, C_a, and the other between the body and the earth, C_b. Then, the resulting capacitance C between the plate and the earth becomes larger by ΔC

$$C = C_1 + \Delta C = C_1 + \frac{C_a C_b}{C_a + C_b}. \tag{6.9}$$

With the appropriate apparatus, this phenomenon can be used for occupancy detection. What is required is to measure a capacitance between a test plate (the probe) and a reference plate (the earth).

[4] At 40 MHz, the dielectric constant of muscle, skin, and blood is about 97. For fat and bone, it is near 15.

[5] Here, by "earth" we mean any large object, such as earth, lake, metal fence, car, ship, airplane, and so forth.

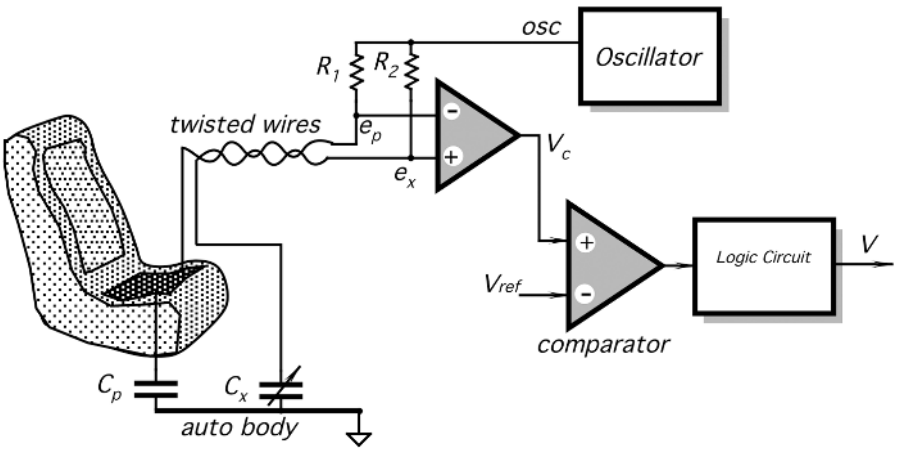


Fig. 6.4. An automotive capacitive intrusion detector.

Figure 6.4 illustrates a capacitive security system for an automobile [3]. A sensing probe is imbedded into a car seat. It can be fabricated as a metal plate, metal net, a conductive fabric, and so forth. The probe forms one plate of a capacitor C_p. The other plate of the capacitor is formed either by the body of an automobile or by a separate plate positioned under a floor mat. A reference capacitor C_x is composed of a simple fixed or trimming capacitor which should be placed close to the seat probe. The probe plate and the reference capacitor are respectively connected to two inputs of a charge detector (resistors R_1 and R_2). The conductors preferably should be twisted to reduce the introduction of spurious signals as much as possible. For instance, strips of twinflex cabling were found to be quite adequate. A differential charge detector is controlled by an oscillator which produces square pulses (Fig. 6.5). Under the no-seat-occupied conditions, the reference capacitor is adjusted to be approximately equal to C_p. Resistors and the corresponding capacitors define time constants of the networks. Both RC circuits have equal time constants τ_1. Voltages across the resistors are fed into the inputs of a differential amplifier, whose output voltage V_c is near zero. Small spikes at the output is the result of some imbalance. When a person is positioned on the seat, his (her) body forms an additional capacitance in parallel with C_p, thus increasing the time constant of the R_1C_p network from τ_1 to τ_2. This is indicated by the increased spike amplitudes at the output of a differential amplifier. The comparator compares V_c with a predetermined threshold voltage V_{ref}. When the spikes exceed the threshold, the comparator sends an indication signal to the logic circuit that generates signal V manifesting the car occupancy. It should be noted that a capacitive detector is an active sensor, because it essentially required an oscillating test signal to measure the capacitance value.

When a capacitive occupancy (proximity) sensor is used near or on a metal device, its sensitivity may be severely reduced due to a capacitive coupling between the electrode and the device's metallic parts. An effective way to reduce that stray

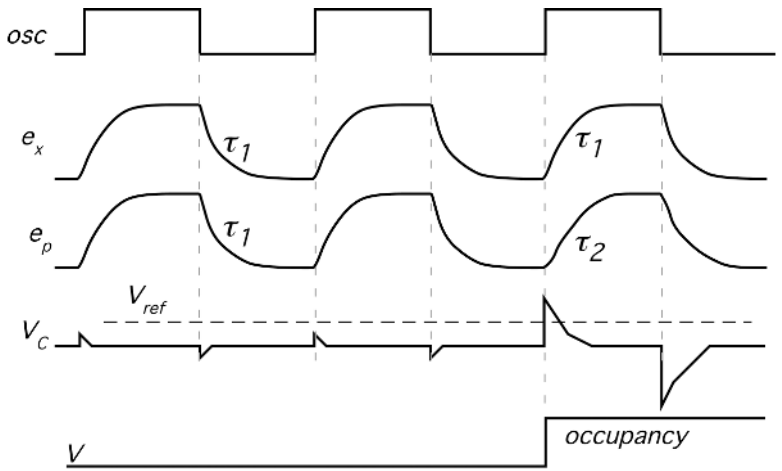


Fig. 6.5. Timing diagrams for a capacitive intrusion detector.

capacitance is to use driven shields. Figure 6.6A shows a robot with a metal arm. The arm moves near people and other potentially conductive objects with which it could collide if the robot's control computer is not provided with advance information on the proximity to the obstacles. An object, while approaching the arm, forms a capacitive coupling with it, which is equal to C_{so}. The arm is covered with an electrically isolated conductive sheath called an *electrode*. As Fig. 6.3 shows, a coupling capacitance can be used to detect the proximity. However, the nearby massive metal arm (Fig. 6.6B) forms a much stronger capacitive coupling with the electrode which drags the electric field from the object. An elegant solution[6] is to shield the electrode from the arm by an intermediate shield, as shown in Fig. 6.6C. The sensor's assembly is a multilayer cover for the robot's arm, where the bottom layer is an insulator, then there is a large electrically conductive shield, then another layer of insulation, and on the top is a narrower sheet of the electrode. To reduce the capacitive coupling between the electrode and the arm, the shield must be at the same potential as the electrode; that is, its voltage must be driven by the electrode voltage (thus, the name is *driven shield*). Hence, there would be no electric field between them; however, there will be a strong electric field between the shield and the arm. The electric field is squeezed out from beneath the electrode and distributed toward the object. Figure 6.7 shows a simplified circuit diagram of a square-wave oscillator whose frequency depends on the net input capacitance, comprised of C_{sg} (sensor-to-ground), C_{so} (sensor-to-object), and C_{og} (object-to-ground). The electrode is connected to the shield through a voltage follower. The frequency-modulated signal is fed into the robot's computer for controlling the arm movement. This arrangement allows us to detect the proximity to conductive objects over the range of 30 cm.

[6] This device was developed for NASA's Jet Propulsion Laboratory by M.S. Katow at Palnning Research Corp.

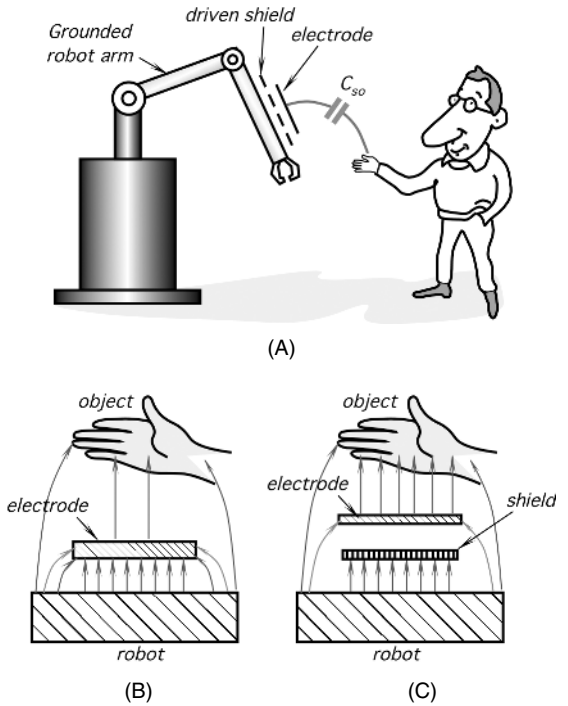


Fig. 6.6. Capacitive proximity sensor. A driven shield is positioned on the metal arm of a grounded robot (A). Without the shield, the electric field is mostly distributed between the electrode and the robot (B), whereas a driven shield directs the electric field from the electrode toward the object (C).

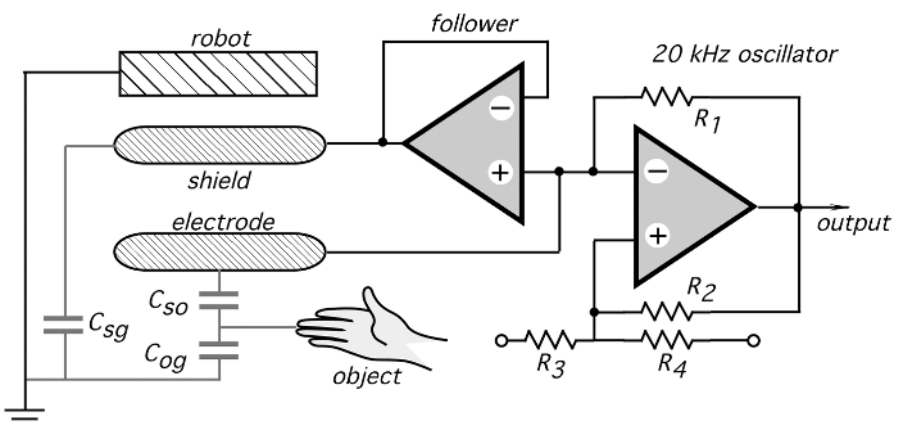


Fig. 6.7. Simplified circuit diagram of a frequency modulator controlled by the input capacitances.

6.4 Triboelectric Detectors

Any object can accumulate static electricity on its surface. These naturally occurring charges arise from the triboelectric effect (i.e., a process of charge separation due to object movements, friction of clothing fibers, air turbulence, atmosphere electricity, etc.) (see Section 3.1 of Chapter 3). Usually, air contains either positive or negative ions that can be attracted to the human body, thus changing its charge. Under the idealized static conditions, an object is not charged: Its bulk charge is equal to zero. In reality, any object which at least temporarily is isolated from the ground can exhibit some degree of its bulk charge imbalance. In other words, it becomes a carrier of electric charges.

Any electronic circuit is made up of conductors and dielectrics. If a circuit is not shielded, all of its components exhibit a certain capacitive coupling to the surrounding objects. In practice, the coupling capacitance may be very small—on the order of 1 pF or less. A pickup *electrode* can be added to the circuit's input to increase its coupling to the environment, very much like in the capacitive detectors discussed in Section 6.3. The electrode can be fabricated in the form of a conductive surface which is well isolated from the ground.

An electric field is established between the surrounding objects and the electrode whenever at least one of them carries electric charges. In other words, all distributed capacitors formed between the electrode and the environmental objects are charged by the static or slow-changing electric fields. Under the no-occupancy conditions, the electric field in the electrode vicinity is either constant or changes relatively slowly.

If a charge carrier (a human or an animal) changes its position (moves away or a new charge carrying an object enters into the vicinity of the electrode), the static electric field is disturbed. This results in a redistribution of charges between the coupling capacitors, including those which are formed between the input electrode and the surroundings. The charge magnitude depends on the atmospheric conditions and the nature of the objects. For instance, a person in dry man-made clothes walking along a carpet carries a million times stronger charge than a wet intruder who has come from the rain. An electronic circuit can be adapted to sense these variable charges at its input. In other words, it can be made capable of converting the induced variable charges into electric signals that may be amplified and further processed. Thus, static electricity, which is a naturally occurring phenomenon, can be utilized to generate alternating signals in the electronic circuit to indicate the movement of objects.

Figure 6.8 shows a monopolar triboelectric motion detector. It is composed of a conductive electrode connected to an analog impedance converter made with a MOS transistor Q_1 a bias resistor R_1, an input capacitance C_0, a gain stage, and a window comparator [4]. Whereas the rest of the electronic circuit may be shielded, the electrode is exposed to the environment and forms a coupling capacitor C_p with the surrounding objects. In Fig. 6.8, static electricity is exemplified by positive charges distributed along the person's body. Being a charge carrier, the person generates an electric field, having intensity E. The field induces a charge of the opposite sign in the electrode. Under static conditions, when the person does not move, the field intensity is constant and the input capacitance C_0 is discharged through a bias resistor R_1. That

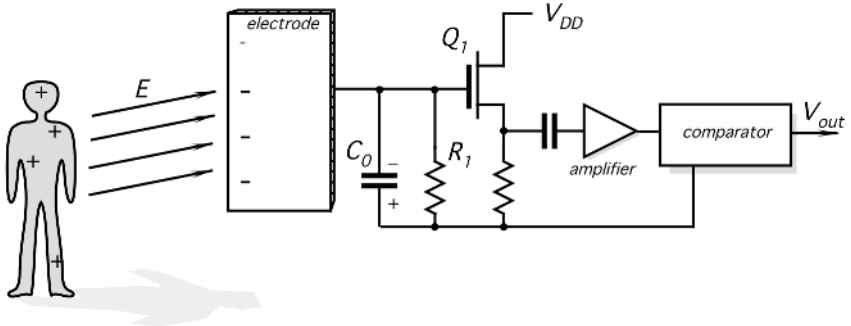


Fig. 6.8. Monopolar triboelectric motion detector.

resistor must be selected of a very high value (on the order of $10^{10}\Omega$ or higher), to make the circuit sensitive to relatively slow motions.

When the person moves, the intensity E of the electric field changes. This induces the electric charge in the input capacitor C_0 and results in the appearance of a variable electric voltage across the bias resistor. That voltage is fed through the coupling capacitor into the gain stage whose output signal is applied to a window comparator. The comparator compares the signal with two thresholds, as is illustrated in a timing diagram in Fig. 6.9B. A positive threshold is normally higher than the baseline static signal, and the other threshold is lower. During human movement, a signal at the comparator's input deflects either upward or downward, crossing one of the thresholds. The output signals from the window comparator are square pulses which can be utilized and further processed by conventional data processing devices. It should be noted that contrary to a capacitive motion detector, which is an active sensor, a triboelectric detector is passive; that is, it does not generate or transmit any signal.

There are several possible sources of interference which may cause spurious detections by the triboelectric detectors; that is, the detector may be subjected to transmitted noise resulting in a false-positive detection. Among the noise sources are 60- or 50-Hz power line signals, electromagnetic fields generated by radio stations, power electric equipment, lightnings, and so forth. Most of these interferences generate electric fields which are distributed around the detector quite uniformly and can be compensated for by employing a symmetrical input circuit with a significant common-mode rejection ratio.

6.5 Optoelectronic Motion Detectors

By far the most popular intrusion sensors are the optoelectronic motion detectors. They rely on electromagnetic radiation in the optical range, specifically having wavelengths from 0.4 to 20 μm. This covers the visible, near-infrared and part of the far-infrared spectral ranges. The detectors are primarily used for the indication of movement of people and animals. These detectors operate over distance ranges up to several hundred meters and, depending on the particular need, may have either a narrow or wide field of view.

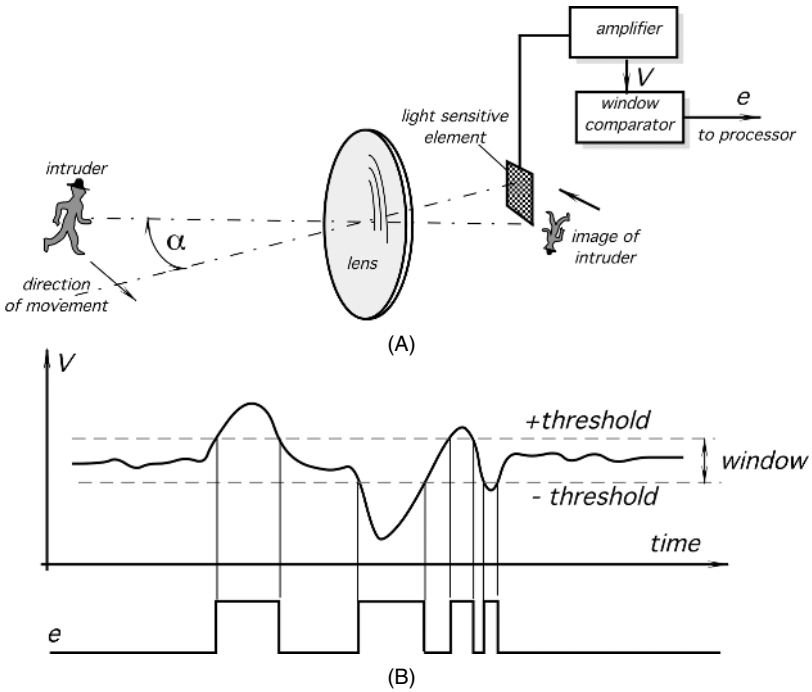


Fig. 6.9. General arrangement of an optoelectronic motion detector. A lens forms an image of a moving object (intruder). When the image crosses the optical axis of the sensor, it superimposes with the sensitive element (A). The element responds with the signal, which is amplified and compared to two thresholds in the window comparator (B).

The operating principle of the optical motion detectors is based on the detection of light (either visible or not) emanated from the surface of a moving object into the surrounding space. Such radiation may be originated either by an external light source and then reflected by the object or it may be produced by the object itself in the form of natural emission. The former case is classified as an active detector and the latter is classified as a passive detector. Hence, an active detector requires an additional light source, (e.g., daylight, electric lamp, an infrared light-emitting diode (LED), and so forth. The passive detectors perceive mid- and far-infrared emission from objects having temperatures that are different from the surroundings. Both types of detector use an optical contrast as a means of object recognition.

First, we must consider the limitations of the optoelectronic detectors as opposed to such devices as microwave or ultrasonic devices. Presently, optoelectronic detectors are used almost exclusively to detect the presence or absence of movement qualitatively rather than quantitatively. In other words, the optoelectronic detectors are very useful for indicating whether an object moves or not; however, they cannot distinguish one moving object from another and they cannot be utilized to accurately measure the distance to a moving object or its velocity. The major application areas for the optoelectronic motion detectors are in security systems (to detect intruders),

in energy management (to turn lights on and off), and in the so-called "smart homes," in which they can control various appliances, such as air conditioners, cooling fans, stereo players, and so forth. They also may be used in robots, toys, and novelty products. The most important advantages of an optoelectronic motion detector are simplicity and low cost.

6.5.1 Sensor Structures

The general structure of an optoelectronic motion detector is shown in Fig. 6.9A. Regardless of what kind of sensing element is employed, the following components are essential: a focusing device (a lens or a focusing mirror), a light-detecting element, and a threshold comparator. An optoelectronic motion detector resembles a photographic camera. Its focusing components create an image of its field of view on a focal plane. Although there is no mechanical shutter like in a camera, a light-sensitive element is used in place of the film. The element converts the focused light into an electric signal.

Let us assume that the motion detector is mounted in a room. A focusing lens creates an image of the room on a focal plane where the light-sensitive element is positioned. If the room is unoccupied, the image is static and the output signal from the element is steady stable. When an "intruder" enters the room and keeps moving, his image on the focal plane also moves. In a certain moment, the intruder's body is displaced by an angle α and the image overlaps with the element. This is an important point to understand: The detection is produced only at the moment when the object's image either coincides with the detector's surface or clears it; that is, no overlapping—no detection. Assuming that the intruder's body creates an image whose electromagnetic flux is different from that of the static surroundings, the light-sensitive element responds with a deflecting voltage V. In other words, to cause detection, a moving image must have a certain degree of optical contrast with its surroundings.

Figure 6.9B shows that the output signal is compared with two thresholds in the window comparator. The purpose of the comparator is to convert the analog signal V into two logic levels: $\emptyset$ = no motion detected and 1 = motion is detected. In most cases, the signal V from the element first must be amplified and conditioned before it becomes suitable for the threshold comparison. The window detector contains both the positive and negative thresholds, whereas the signal V is positioned in between. Whenever the image of a moving object overlaps with the light-sensitive element, the voltage V deflects from its baseline position and crosses one of two thresholds. The comparator generates a positive voltage (1), thus indicating a detection of movement in the field of view. The operation of this circuit is identical to the threshold circuits described earlier for other types of occupancy detector.

It may be noted from Fig. 6.9 that the detector has quite a narrow field of view: If the intruder keeps moving, his image will overlap with the sensor only once; after that, the window comparator output will a produce steady $\emptyset$. This is a result of the small area of the sensing element. In some instances, when a narrow field of view is

required, it is quite all right; however, in the majority of cases, a much wider field of view is desirable. This can be achieved by several methods described in the following subsections.

6.5.1.1 Multiple Sensors

An array of detectors may be placed in the focal plane of a focusing mirror or lens. Each individual detector covers a narrow field of view; however, in combination, they protect larger area. All detectors in the array either must be multiplexed or otherwise interconnected to produce a combined detection signal.

6.5.1.2 Complex Sensor Shape

If the detector's surface area is sufficiently large to cover an entire angle of view, it may be optically broken into smaller elements, thus creating an equivalent of a multiple-detector array. To break the surface area into several parts, one may shape the sensing element in an odd pattern like that shown in Fig. 6.10A. Each part of the element acts as a separate light detector. All such detectors are electrically connected either in parallel or in series while being arranged in a serpentine pattern. The parallel or serially connected detectors generate a combined output signal, (e.g., voltage v), when the image of the object moves along the element surface crossing sensitive and nonsensitive areas alternatively. This results in an alternate signal v at the detector terminals. Each sensitive and nonsensitive area must be sufficiently large to overlap with most of the object's image.

6.5.1.3 Image Distortion

Instead of making the detector in a complex shape, an image of an entire field of view may be broken into several parts. This can be done by placing a distortion mask in front of the detector having a sufficiently large area, as is depicted in Fig. 6.10B. The mask is opaque and allows the formation of an image on the detector's surface only

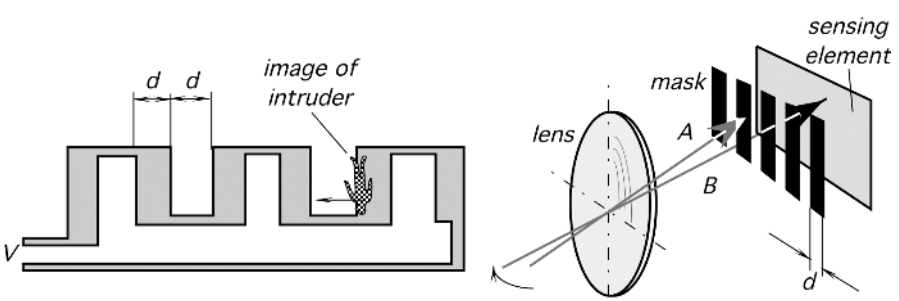


Fig. 6.10. Complex shape of a sensing element (A); an image-distortion mask (B).

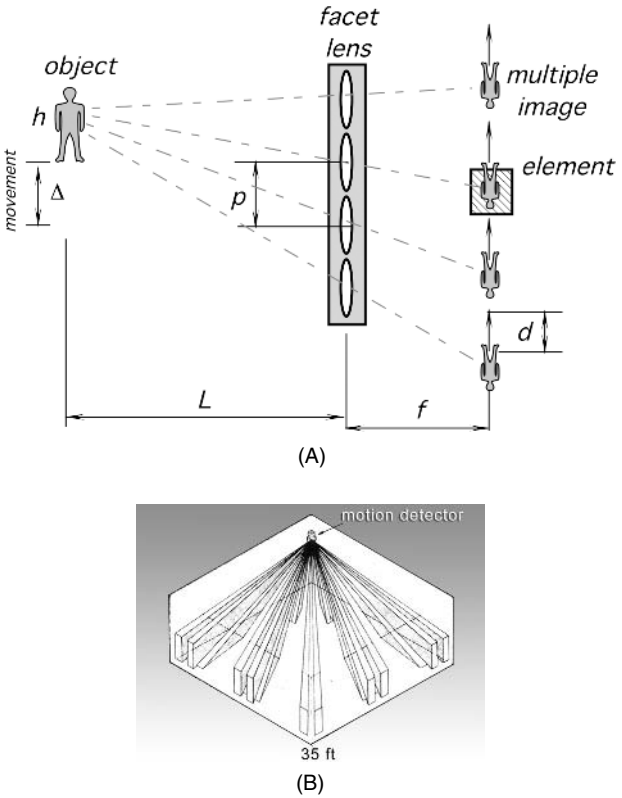


Fig. 6.11. A facet lens creates multiple images near the sensing element (A); sensitive zones created by a complex facet lens (B).

within its clearings. The mask operation is analogous to the complex sensor's shape as described in Section 6.5.1.2.

6.5.1.4 Facet Focusing Element

Another way of broadening the field of view while employing a small-area detector is to use multiple focusing devices. A focusing mirror or a lens may be divided into an array of smaller mirrors or lenses called facets. Each facet creates its own image, resulting in multiple images as shown in Fig. 6.11A. When the object moves, the images also move across the element, resulting in an alternate signal. By combining multiple facets, it is possible to create any desirable detecting pattern in the field of view, in both horizontal and vertical planes. Positioning of the facet lens, focal distances, number, and the pitch of the facets (the distance between the optical axes of two adjacent facets) may by calculated in every case by applying rules of geometrical optics. The following practical formulas may be applied to find the focal length of a facet:

$$f = \frac{L\,d}{\Delta}, \tag{6.10}$$

and the facet pitch is

$$p = 2nd, \tag{6.11}$$

where L is the distance to the object, d is the width of the sensing element, n is the number of sensing elements (evenly spaced), and Δ is the object's minimum displacement which must result in detection. For example, if the sensor has two sensing elements of $d = 1$ mm, each of which are positioned at 1 mm apart, and the object's minimum displacement $\Delta = 25$ cm at a distance $L = 10$ m, the facet focal length is calculated from Eq. (6.10) as $f = (1000\text{ cm})(0.1\text{ cm})/25\text{ cm} = 4$ cm, and the facets should be positioned with a pitch of $p = 8$ mm from one another as per Eq. (6.11).

By combining facets, one may design a lens which covers a large field of view (Fig. 6.11B) where each facet creates a relatively narrow-angle sensitive zone. Each zone projects an image of an object into the same sensing element. When the object moves, it crosses the zone boundaries, thus modulating the sensor's output.

6.5.2 Visible and Near-Infrared Light Motion Detectors

Most of the objects (apart from those very hot) radiate electromagnetic waves only in a far-infrared spectral range. Hence, visible and near-infrared light motion detectors have to rely on the additional source of light which illuminates the object. The light is reflected by the object's body toward the focusing device for the subsequent detection. Such illumination may be sunlight or the invisible infrared light from an additional near-infrared light source (a projector). The use of a visible light for detecting moving objects goes back to 1932 when, in the preradar era, inventors were looking for ways to detect flying airplanes. In one invention, an airplane detector was built in the form of a photographic camera where the focusing lens made of glass was aimed at the sky. A moving plane's image was focused on a selenium photodetector, which reacted to the changing contrast in the sky image. Naturally, such a detector could operate only in daytime to detect planes flying below clouds. Obviously, those detectors were not very practical. Another version of a visible light motion detector was patented for less demanding applications: controlling lights in a room [5] and making interactive toys [6].

To turn the lights off in an unoccupied room, the visible-range motion detector (Motion Switch manufactured by Intermatic, Inc., IL) was combined with a timer and a power solid-state relay. The detector is activated when the room is illuminated. Visible light carries a relatively high energy and may be detected by quantum photovoltaic or photoconductive cells whose detectivity is quite high. Thus, the optical system may be substantially simplified. In the Motion Switch, the focusing device was built in a form of a pinhole lens (Fig. 6.12C). Such a lens is just a tiny hole in an opaque foil. To avoid a light-wave diffraction, the hole diameter must be substantially larger than the longest detectable wavelength. Practically, the Motion Switch has a three-facet pinhole lens, where each hole has an aperture of 0.2 mm (Fig. 6.12C). Such a lens has a theoretically infinitely deep focusing range; hence, the photodetector can be positioned at any distance from it. For practical reasons, that distance was calculated for a maximum of the object's displacement and the photoresistor's dimensions used in the design. The photoresistor was selected with a serpentine pattern of the sensing

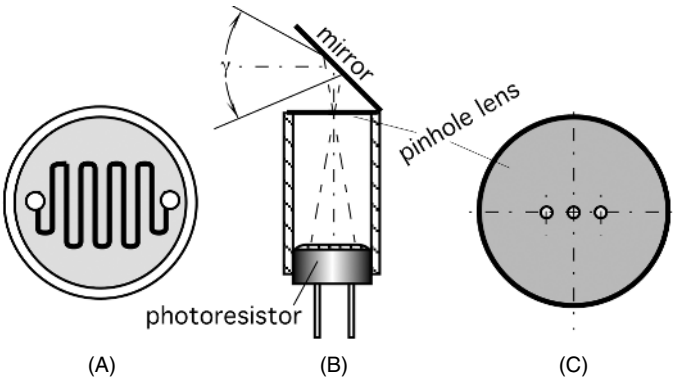


Fig. 6.12. A simple optical motion detector for a light switch and toys: (A) the sensitive surface of a photoresistor forms a complex sensing element; (B) a flat mirror and a pinhole lens form an image on a surface of the photoresistor; (C) a pinhole lens.

element (Fig. 6.12A) and connected into a circuit which responds only to its variable (alternate) component. When the room is illuminated, the sensor acts as a miniature photographic camera: an image of its field of view is formed on a surface of the photoresistor. Moving people in the room cause the image to change in such a way that the optical contrast changes across the serpentine pattern of the photoresistor. In turn, its resistive value changes, which results in the modulation of the electric current passing through the element. This signal is further amplified and compared with a predetermined threshold. Upon crossing that threshold, the comparator generates electric pulses which reset a 15-min timer. If no motion is detected within 15 min from the last movement, the timer turns lights off in the room. Then, it may be turned on again only manually, because this motion detector does not function in darkness.

6.5.3 Far-Infrared Motion Detectors

Another version of a motion detector operates in the optical range of thermal radiation, the other name for which is far infrared. Such detectors are responsive to radiative-heat exchange between the sensing element and the moving object [7–9]. Here, we will discuss the detection of moving people; however, the technique which is described may be modified for other warm or cold objects.

The principle of thermal motion detection is based on the physical theory of the emission of electromagnetic radiation from any object whose temperature is above absolute zero. The fundamentals of this theory are described in Section 3.12.3 of Chapter 3. We recommend that the reader first become familiar with that section before going further.

For motion detection, it is essential that the surface temperature of an object be different from that of the surrounding objects, so that a thermal contrast would exist. All objects emanate thermal radiation from their surfaces and the intensity of that radiation is governed by the Stefan–Boltzmann law [Eq. (3.133)]. If the object is

warmer than the surroundings, its thermal radiation is shifted toward shorter wavelengths and its intensity becomes stronger. Many objects whose movement is to be detected are nonmetals; hence, they radiate thermal energy quite uniformly within a hemisphere (Fig. 3.45A of Chapter 3). Moreover, the dielectric objects generally have a high emissivity. Human skin is one of the best emitters, with an emissivity of over 90% (See Table A.18), whereas most fabrics have also high emissivities between 0.74 and 0.95. In the following subsections, we describe two types of far-infrared motion detectors. The first utilizes a passive infrared (PIR) sensor and the second has active far-infrared (AFIR) elements.

6.5.3.1 PIR Motion Detectors

These detectors became extremely popular for the security and energy management systems. The PIR sensing element must be responsive to far-infrared radiation within a spectral range from 4 to 20 μm, where most of the thermal power emanated by humans is concentrated. There are three types of sensing element which are potentially useful for that detector: thermistors, thermopiles, and pyroelectrics However, the pyroelectric elements are used almost exclusively for the motion detection thanks to their simplicity, low cost, high responsivity, and a broad dynamic range. A pyroelectric effect is described in Section 3.7 of Chapter 3 and some detectors are covered in Section 14.6.3 of Chapter 14. Here, we will see how that effect may be used in a practical sensor design.

A pyroelectric material generates an electric charge in response to thermal energy flow through its body. In a very simplified way, it may be described as a secondary effect of a thermal expansion (Fig. 6.13). Because all pyroelectrics are also piezoelectrics, the absorbed heat causes the front side of the sensing element to expand. The resulting thermally induced stress leads to the development of a piezoelectric charge on the element electrodes. This charge is manifested as voltage across the electrodes deposited on the opposite sides of the material. Unfortunately, the piezoelectric properties of the element have also a negative effect. If the sensor is subjected to a minute mechanical stress due to any external force, it also generates a charge which, in most cases, is indistinguishable from that caused by the infrared heat waves.

To separate thermally induced charges from the piezoelectrically induced charges, a pyroelectric sensor is usually fabricated in a symmetrical form (Fig. 6.14A). Two

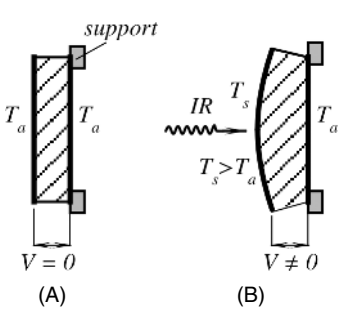


Fig. 6.13. A simplified model of a pyroelectric effect as a secondary effect of piezoelectricity. Initially, the element has a uniform temperature (A); upon exposure to thermal radiation, its front side expands, causing a stress-induced charge (B).

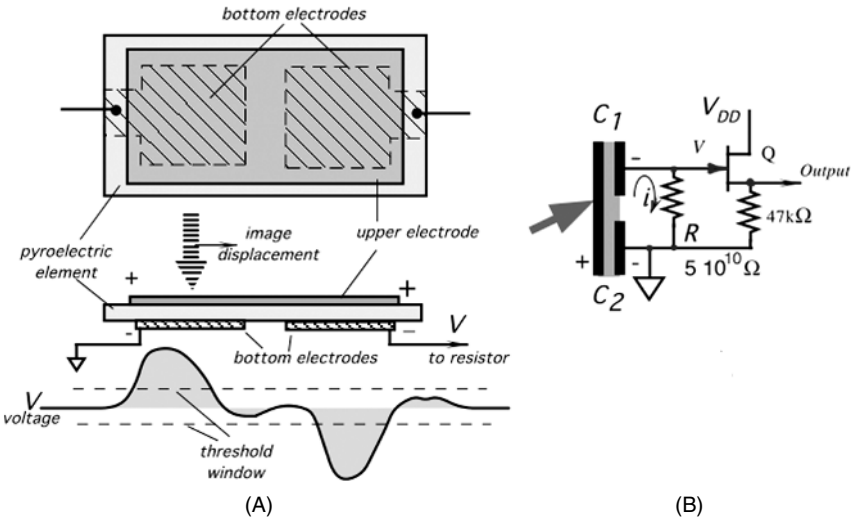


Fig. 6.14. Dual pyroelectric sensor. (A) A sensing element with a front (upper) electrode and two bottom electrodes deposited on a common crystalline substrate. A moving thermal image travels from left part of the sensor to the right, generating an alternate voltage across bias resistor, *R* (B).

identical elements are positioned inside the sensor's housing. The elements are connected to the electronic circuit in such a manner as to produce the out-of-phase signals when subjected to the same in-phase inputs. The idea is that interferences produced, say, by the piezoelectric effect or spurious heat signals are applied to both electrodes simultaneously (in phase) and, thus, will be canceled at the input of the circuit, whereas the variable thermal radiation to be detected will be absorbed by only one element at a time, thus avoiding a cancellation.

One way to fabricate a symmetrical sensor is to deposit two pairs of electrodes on both sides of a pyroelectric element. Each pair forms a capacitor which may be charged either by heat or by a mechanical stress. The electrodes on the upper side of the sensor are connected together, forming one continuous electrode, whereas the two bottom electrodes are separated, thus creating the opposite, serially connected capacitors. Depending on the side where the electrodes are positioned, the output signal will have either a positive or negative polarity for the thermal influx. In some applications, a more complex pattern of the sensing electrodes may be required (e.g., to form predetermined detection zones), so that more than one pair of electrodes is needed. In such a case, for better rejection of the in-phase signals (common-mode rejection), the sensor still should have an even number of pairs, where positions of the pairs alternate for better geometrical symmetry. Sometimes, such an alternating connection is called an interdigitized electrode.

A symmetrical sensing element should be mounted in such a way as to assure that both parts of the sensor generate the same signal if subjected to the same external factors. At any moment, the optical component must focus a thermal image of an

object on the surface of one part of the sensor only, which is occupied by a single pair of electrodes. The element generates a charge only across the electrode pair subjected to a heat flux. When the thermal image moves from one electrode to another, the current i flowing from the sensing element to the bias resistor R (Fig. 6.14B) changes form zero, to positive, then to zero, to negative, and again to zero (Fig. 6.14A lower portion). A JFET transistor Q is used as an impedance converter. The resistor R value must be very high. For example, a typical alternate current generated by the element in response to a moving person is on the order of 1 pA (10^{-12} A). If a desirable output voltage for a specific distance is $v = 50$ mV, according to Ohm's law the resistor value is $R = v/i = 50$ GΩ ($5 \times 10^{10}\Omega$). Such a resistor can not be directly connected to a regular electronic circuit; hence, transistor Q serves as a voltage follower (the gain is close to unity). Its typical output impedance is on the order of several kilohms.

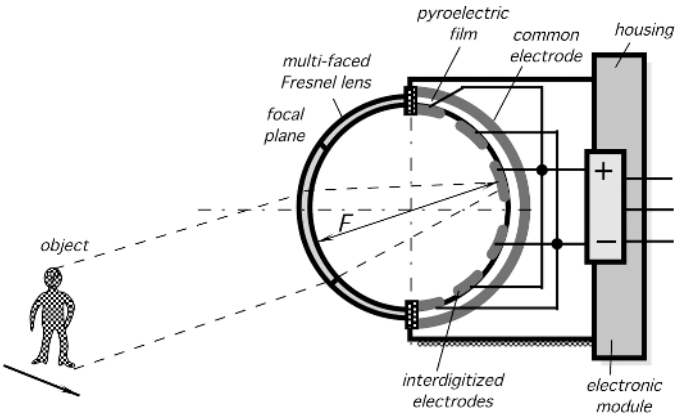
Table A.9 lists several crystalline materials which possess a pyroelectric effect and can be used for the fabrication of sensing elements. The most often used are the ceramic elements, thanks to their low cost and ease of fabrication. The pyroelectric coefficient of ceramics to some degree may be controlled by varying their porosity (creating voids inside the sensor's body). An interesting pyroelectric material is a polymer film polyvinylidene fluoride (PVDF) which, although not as sensitive as most of the solid-state crystals, has the advantages of being flexible and inexpensive. In addition, it can be produced in any size and may be bent or folded in any desirable fashion.

In addition to the sensing element, an infrared motion detector needs a focusing device. Some detectors employ parabolic mirrors, but the Fresnel plastic lenses (Section 4.6 of Chapter 4) become more and more popular because they are inexpensive, may be curved to any desirable shape, and, in addition to focusing, act as windows, protecting the interior of the detector from outside moisture and pollutants.

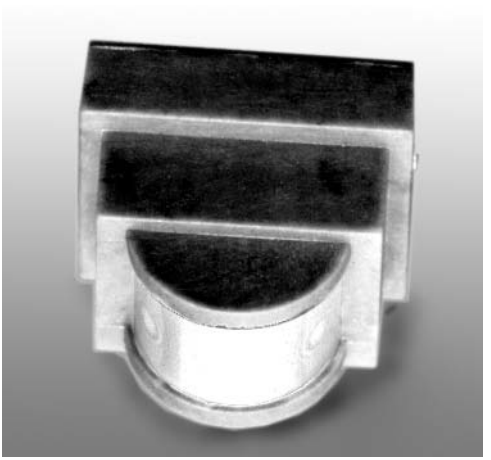
To illustrate how a plastic Fresnel lens and a PVDF film can work together, let us look at the motion detector depicted in Fig. 6.15A. It uses a polyethylene multifaceted curved lens and a curved PVDF film sensor [7]. The sensor design combines two methods described earlier: a facet lens and a complex electrode shape. The lens and the film are curved with the same radii of curvature equal to one-half of the focal distance f, thus assuring that the film is always positioned in the focal plane of the corresponding facet of the lens. The film has a pair of large interdigitized electrodes which are connected to the positive and negative inputs of a differential amplifier located in the electronic module. The amplifier rejects common-mode interference and amplifies a thermally induced voltage. The side of the film facing the lens is coated with an organic coating to improve its absorptivity in the far-infrared spectral range. This design results in a fine resolution (detection of small displacement at a longer distance) and a very small volume of the detector (Fig. 6.15B). Small detectors are especially useful for the installation in devices where overall dimensions are critical. For instance, one application is a light switch where the detector must be mounted into the wall plate of a switch.

6.5.3.2 PIR Sensor Efficiency Analysis

Regardless of the type of optical device employed, all modern PIR detectors operate on the same physical effect—pyroelectricity. To analyze the performance of such a



(A)



(B)

Fig. 6.15. Far-infrared motion detector with a curved Fresnel lens and a pyroelectric PVDF film: (A) internal structure of the sensor; (B) external appearance of the sensor.

sensor, first we must calculate the infrared power (flux), which is converted into an electric charge by the sensing element. The optical device focuses thermal radiation into a miniature thermal image on the surface of the sensor. The energy of an image is absorbed by the sensing element and is converted into heat. That heat, in turn, is converted by the pyroelectric crystalline element into a minute electric current.

To estimate a power level at the sensor's surface, let us make some assumptions. We assume that the moving object is a person whose effective surface area is b (Fig. 6.16) and the temperature along this surface (T_b) is distributed uniformly and is expressed in Kelvin. The object is assumed to be a diffuse emitter (radiates uniformly within the hemisphere having a surface area of $A = 2\pi L^2$). Also, we assume that the focusing device makes a sharp image of an object at any distance. For this calculation,

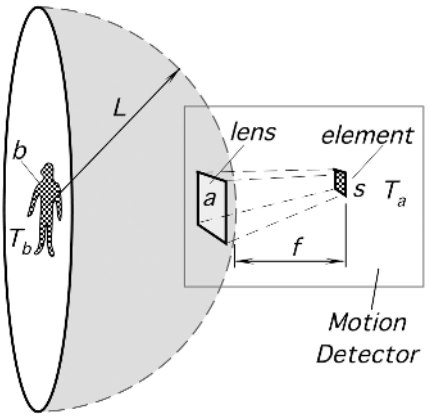


Fig. 6.16. Formation of a thermal image on the sensing element of the PIR motion detector.

we select a lens which has a surface area a. The sensor's temperature (in K) is T_a, the same as that of ambient.

The total infrared power (flux) lost to surroundings from the object can be determined from the Stefan–Boltzmann law:

$$\Phi = b\varepsilon_a\varepsilon_b\sigma\left(T_b^4 - T_a^4\right), \tag{6.12}$$

where σ is the Stefan–Boltzmann constant and ε_b and ε_a are the object and the surrounding emissivities, respectively. If the object is warmer than the surroundings (which is usually the case), this net infrared power is distributed toward an open space. Because the object is a diffusive emitter, we may consider that the same flux density may be detected along an equidistant surface. In other words, the intensity of infrared power is distributed uniformly along the spherical surface having radius L.

Assuming that the surroundings and the object's surface are ideal emitters and absorbers ($\varepsilon_b = \varepsilon_a = 1$) and the sensing element's emissivity is ε_s, the net radiative flux density at distance L can be derived as

$$\phi = \frac{b}{2\pi L^2}\varepsilon_s\sigma\left(T_b^4 - T_a^4\right). \tag{6.13}$$

The lens efficiency (transmission coefficient) is γ, which theoretically may vary from 0 to 0.92 depending on the properties of the lens material and the lens design. For the polyethylene Fresnel lenses, its value is in the range from 0.4 to 0.7. After ignoring a minor nonlinearity related to the fourth power of temperatures in Eq. (6.13), thermal power absorbed by the element can be expressed as

$$\Phi_s = a\gamma\phi \approx \frac{2\sigma\varepsilon_s}{\pi L^2}a\gamma T_a^3\left(T_b - T_a\right). \tag{6.14}$$

surface of the sensing element is inversely proportional to the squared distance from the object and directly proportional to the areas of the lens and the object. It is important to note that in the case of a multifacet lens, the lens area a relates only to a single facet and not to the total lens area. If the object is warmer than the sensor, the flux Φ_s is positive. If the object is cooler, the flux becomes negative, meaning it changes its direction: The heat goes from the sensor to the object. In reality, this may

happen when a person walks into a warm room from the cold outside. The surface of his clothing will be cooler than the sensor and the flux will be negative. In the following discussion, we will consider that the object is warmer than the sensor and the flux is positive.

A maximum operating distance for given conditions can be determined by the noise level of the detector. For reliable discrimination, the worst-case noise power must be at least three to five times smaller than that of the signal.

The pyroelectric sensor is a converter of thermal energy flow into electric charge. The energy flow essentially demands the presence of a thermal gradient across the sensing element. In the detector, the element of thickness h has the front side exposed to the lens, and the opposite side faces the detector's interior housing, which normally is at ambient temperature T_a. The front side of the sensor element is covered with a heat-absorbing coating to increase its emissivity ε_s to the highest possible level, preferably close to unity. When thermal flux Φ_s is absorbed by the element's front side, the temperature increases and heat starts propagating through the sensor toward its rear side. Because of the pyroelectric properties, electric charge is developing on the element surfaces in response to the heat flow.

Upon influx of the infrared radiation, the temperature of the sensor element increases (or decreases) with the rate, which can be derived from the absorbed thermal power Φ_s and thermal capacity C of the element:

$$\frac{dT}{dt} \approx \frac{\Phi_s}{C}, \tag{6.15}$$

where t is time. This equation is valid during a relatively short interval (immediately after the sensor is exposed to the thermal flux) and can be used to evaluate the signal magnitude. The electric current generated by the sensor can be found from the fundamental formula

$$i = \frac{dQ}{dt}, \tag{6.16}$$

where Q is the electric charge developed by the pyroelectric sensor. This charge depends on the sensor's pyroelectric coefficient P, the sensor's area s, and the temperature change dT:

$$dQ = Ps\,dT. \tag{6.17}$$

Thermal capacity C can be derived through a specific heat c of the material, area s, and thickness of the element h:

$$C = csh. \tag{6.18}$$

By substituting Eqs. (6.15), (6.17), and (6.18) into Eq. (6.16), we can evaluate the peak current which is generated by the sensor in response to the incident thermal flux:

$$i = \frac{Ps\,dT}{dt} = \frac{Ps\Phi_s}{csh} = \frac{P}{hc}\Phi_s. \tag{6.19}$$

To establish relationship between the current and the moving object, the flux from Eq. (6.14) has to be substituted into Eq. (6.19):

$$i = \frac{2Pa\sigma\gamma}{\pi hc} b T_a^3 \frac{\Delta T}{L^2}, \tag{6.20}$$

where $\Delta T = (T_b - T_a)$.

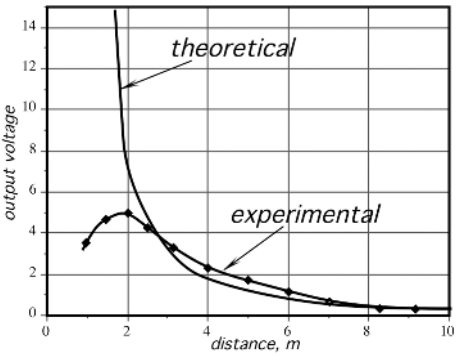


Fig. 6.17. Calculated and experimental amplitudes of output signals in a PIR detector.

There are several conclusions which can be drawn from Eq. (6.20). The first part of the equation (the first ratio) characterizes a detector and the rest relates to an object. The pyroelectric current i is directly proportional to the temperature difference (thermal contrast) between the object and its surroundings. It is also proportional to the surface area of the object which faces the detector. A contribution of the ambient temperature T_a is not as strong as it might appear from its third power. The ambient temperature must be entered in Kelvin; hence, its variations become relatively small with respect to the scale. The thinner the sensing element, the more sensitive is the detector. The lens area also directly affects signal magnitude. On the other hand, pyroelectric current does not depend on the sensor's area as long as the lens focuses the entire image on a sensing element.

To evaluate Eq. (6.20) further, let us calculate the voltage across the bias resistor. That voltage can be used as an indication of motion. We select a pyroelectric PVDF film sensor with typical properties: $P = 25$ μC/K m^2, $c = 2.4 \times 10^6$ J/m^3 K, $h = 25$ μm, lens area $a = 1$ cm^2, $\gamma = 0.6$, and the bias resistor $R = 10^9 \Omega$ (1 GΩ). We will assume that the object's surface temperature is 27°C and the surface area $b = 0.1$ m2. The ambient temperature $t_a = 20$°C. The output voltage is calculated from Eq. (6.20) as a function of distance L from the detector to the object and is shown in Fig. 6.17.

A graph for Fig. 6.17 was calculated under the assumption that the optical system provides a sharp image at all distances and that the image is no larger that the sensing element area. In practice, this is not always true, especially at shorter distances, where the image is not only out of focus but also may overlap the out-of-phase parts of a symmetrical sensor. The reduction in the signal amplitude at shorter distances becomes apparent: The voltage does not go as high as in the calculated curve.

References

1. Blumenkrantz, S. *Personal and Organizational Security Handbook*. Government Data Publications, Washington, DC, 1989.
2. Ryser, P. and Pfister, G. Optical fire and security technology: Sensor principles and detection intelligence. In: *Transducers'91. International conference on Solid-*

State Sensors and Actuators. Digest of Technical Papers. IEEE, New York, 1991, pp. 579–583.
3. Long, D.J. Occupancy detector apparatus for automotive safety system. U.S. patent 3,898,472, 1975.
4. Fraden, J. Apparatus and method for detecting movement of an object, U.S. patent 5,019,804, 1991.
5. Fraden, J. Motion discontinuance detection system and method. U.S. patent 4,450,351, 1984.
6. Fraden, J. Toy including motion-detecting means for activating same. U.S. patent 4,479,329, 1984.
7. Fraden, J. Motion detector. U.S. patent 4,769,545,1988.
8. Fraden, J. Active infrared motion detector and method for detecting movement. U.S. patent 4,896,039, 1990.
9. Fraden, J. Active far infrared detectors. In: *Temperature. Its Measurement and Control in Science and Industry*. American Institute of Physics, New York, 1992, Vol. 6, Part 2, pp. 831–836.

7

Position, Displacement, and Level

"…If you keep moving in that direction,
we always can displace you to such a position
that is not embarrassing to the level of your wisdom".

—Julius Caesar to his senator

The measurement of position and displacement of physical objects is essential for many applications: process feedback control, performance evaluation, transportation traffic control, robotics, and security systems—just to name the few. By *position*, we mean the determination of the object's coordinates (linear or angular) with respect to a selected reference. *Displacement* means moving from one position to another for a specific distance or angle. In other words, a displacement is measured when an object is referenced to its own prior position rather than to another reference.

A critical distance is measured by *proximity* sensors. In effect, a proximity sensor is a threshold version of a position detector. A position sensor is often a linear device whose output signal represents a distance to the object from a certain reference point. A proximity sensor, however, is a somewhat simpler device which generates the output signal when a certain distance to the object becomes essential for an indication. For instance, many moving mechanisms in process control and robotics use a very simple but highly reliable proximity sensor—the end switch. It is an electrical switch having normally open or normally closed contacts. When a moving object activates the switch by physical contact, the latter sends a signal to a control circuit. The signal is an indication that the object has reached the end position (where the switch is positioned). Obviously, such contact switches have many drawbacks, (e.g., a high mechanical load on a moving object and a hysteresis).

A displacement sensor often is part of a more complex sensor where the detection of movement is one of several steps in a signal conversion (see Fig. 1.1 of Chapter 1). An example is a pressure sensor where pressure is translated into a displacement of a diaphragm, and the diaphragm displacement is subsequently converted into an electrical signal representing pressure. Therefore, the positions sensors, some of which

are described in this chapter, are essential for designs of many other sensors, which are covered in the following chapters of this book.

Position and displacement sensors are static devices whose speed response usually is not critical for the performance.[1] In this chapter, we do not cover any sensors whose response is a function of time, which, by definition, are dynamic sensors. They are covered elsewhere in this book.

When designing or selecting position and displacement detectors, the following questions should be answered:

1. How large is the displacement and of what type (linear, circular)?
2. What resolution and accuracy are required?
3. What is the measured object made of (metal, plastic, fluid, ferromagnetic, etc.)?
4. How much space is available for mounting the detector?
5. How much play is there in the moving assembly and what is the required detection range?
6. What are the environmental conditions (humidity, temperature, sources of interference, vibration, corrosive materials, etc.)?
7. How much power is available for the sensor?
8. How much mechanical wear can be expected over the lifetime of the machine?
9. What is the production quantity of the sensing assembly (limited number, medium volume, mass production)?
10. What is the target cost of the detecting assembly?

A careful analysis will pay big dividends in the long term.

7.1 Potentiometric Sensors

A position or displacement transducer may be built with a linear or rotary *potentiometer* or a *pot* for short. The operating principle of this sensor is based on Eq. (3.54) of Chapter 3 for wire resistance. From the formula, it follows that the resistance linearly relates to the wire length. Thus, by making an object to control the length of the wire, as it is done in a pot, a displacement measurement can be performed. Because a resistance measurement requires passage of an electric current through the pot wire, the potentiometric transducer is of an active type; that is, it requires an excitation signal, (e.g., dc current). A stimulus (displacement) is coupled to the pot wiper, whose movement causes the resistance change (Fig. 7.1A). In most practical circuits, the resistance measurement is replaced by a measurement of voltage drop. The voltage across the wiper of a linear pot is proportional to the displacement d:

$$V = E\frac{d}{D}, \tag{7.1}$$

where D is the full-scale displacement and E is the voltage across the pot (excitation signal). This assumes that there is no loading effect from the interface circuit. If there is an appreciable load, the linear relationship between the wiper position and the output voltage will not hold. In addition, the output signal is proportional to the excitation voltage applied across the sensor. This voltage, if not maintained constant, may be a

[1] Nevertheless, the maximum rate of response is usually specified by the manufacturer.

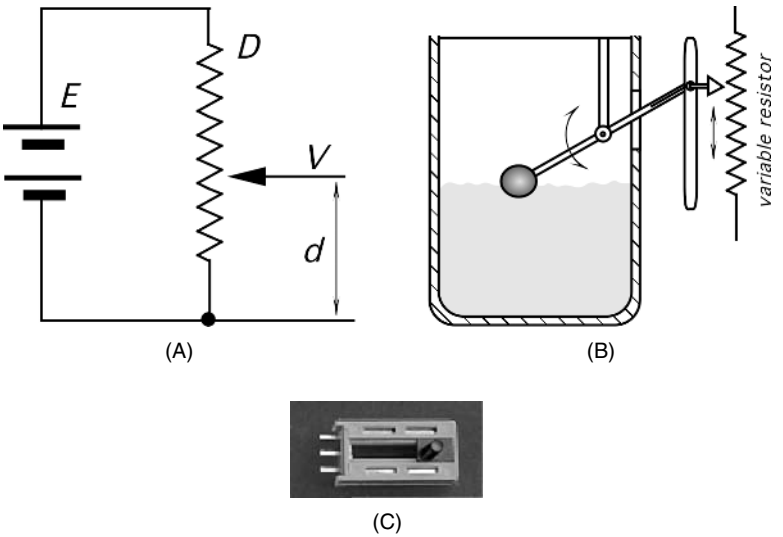


Fig. 7.1. (A) Potentiometer as a position sensor; (B) gravitational fluid level sensor with a float; (C) linear potentiometer. (Courtesy of Piher Group, Tudela, Spain.)

source of error. It should be noted that a potentiometric sensor is a ratiometric device (see Chapter 4); hence the resistance of the pot is not a part of the equation. This means that its stability (e.g., over a temperature range) virtually has no effect on accuracy. For the low-power applications, high-impedance pots are desirable; however, the loading effect must be always considered. Thus, a good voltage follower is required. The wiper of the pot is usually electrically isolated from the sensing shaft.

Figure 7.2A shows one problem associated with a wire-wound potentiometer. The wiper may, while moving across the winding, make contact with either one or two wires, thus resulting in uneven voltage steps (Fig. 7.2B) or a variable resolution. Therefore, when the coil potentiometer with N turns is used, only the average resolution n should be considered:

$$n = \frac{100}{N\%}. \tag{7.2}$$

The force which is required to move the wiper comes from the measured object, and the resulting energy is dissipated in the form of heat. Wire-wound potentiometers are fabricated with thin wires having a diameter on the order of 0.01 mm. A good coil potentiometer can provide an average resolution of about 0.1% of FS (full scale), whereas the high-quality resistive film potentiometers may yield an infinitesimal resolution which is limited only by the uniformity of the resistive material and noise floor of the interface circuit. The continuous-resolution pots are fabricated with conductive plastic, carbon film, metal film, or a ceramic–metal mix which is known as *cermet*. The wiper of the precision potentiometers are made from precious metal alloys. Displacements sensed by the angular potentiometers range from approximately 10° to

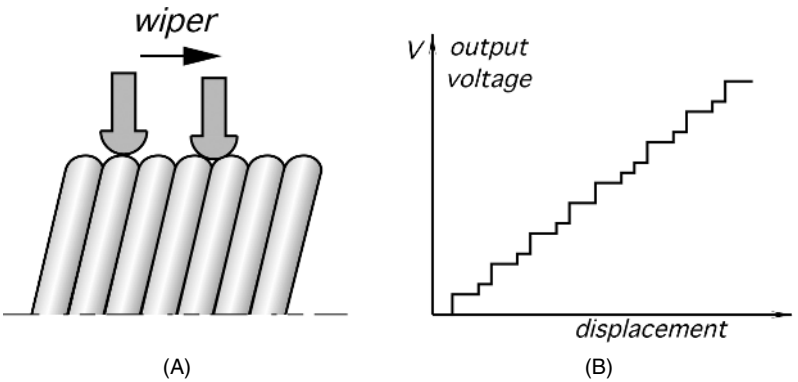


Fig. 7.2. Uncertainty caused by a wire-wound potentiometer: (A) a wiper may contact one or two wires at a time; (B) uneven voltage steps.

over 3000° for the multiturn pots (with gear mechanisms). Although quite useful in some applications, potentiometers have several drawbacks:

1. Noticeable mechanical load (friction)
2. Need for a physical coupling with the object
3. Low speed
4. Friction and excitation voltage cause heating of the potentiometer
5. Low environmental stability

7.2 Gravitational Sensors

A well-known, popular gravitational-level transducer is used in a toilet tank. The transducer's main element is a float—a device whose density is lower than that of water. In most tanks, it is directly coupled to a water valve to keep it either open or shut, depending on how much water the tank holds. The float is a detector of the position of the water surface. For the measurement purposes, the float can be coupled to a position transducer, such as a potentiometric, magnetic, capacitive, or any other direct sensor (Fig. 7.1B). It should be noted that the gravitational sensor is susceptible to various interfering forces, resulting from friction and acceleration. Obviously, such a sensor will not work whenever gravity is altered or absent. A space station or a jet is not an appropriate place for such a sensor.

Inclination detectors, which measure the angle from the direction to the Earth's center of gravity, are employed in road construction, machine tools, inertial navigation systems, and other applications requiring a gravity reference. An old and still quite popular detector of a position is a mercury switch (Figs. 7.3A and 7.3B). The switch is made of a nonconductive (often glass) tube having two electrical contacts and a drop of mercury. When the sensor is positioned with respect to the gravity force in such a way that the mercury moves away from the contacts, the switch is open. A change in the switch orientation causes the mercury to move to the contacts and touch both of

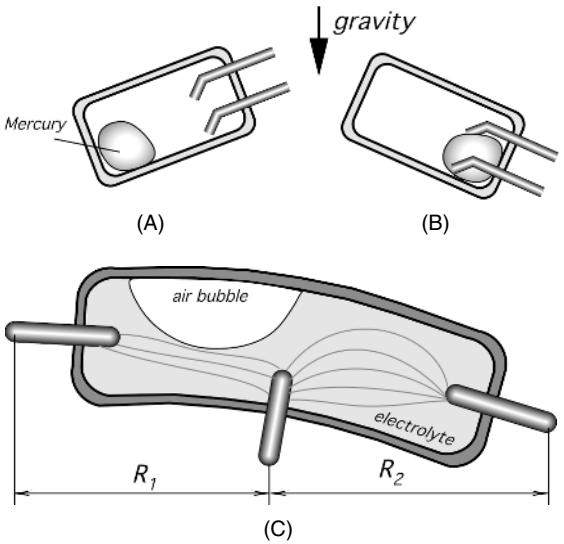


Fig. 7.3. Conductive gravitational sensors: (A) mercury switch in the open position; (B) mercury switch in the closed position; (C) electrolytic tilt sensor.

them, thus closing the switch. One popular application of this design is in a household thermostat, in which the mercury switch is mounted on a bimetal coil which serves as an ambient-temperature sensor. Winding or unwinding the coil in response to room temperature affects the switch's orientation. Opening and closing the switch controls a heating/cooling system. An obvious limitation of this design is its an on–off operation (a bang-bang controller in the engineering jargon). A mercury switch is a threshold device, which snaps when its rotation angle exceeds a predetermined value.

To measure angular displacement with higher resolution, a more complex sensor is required. One elegant design is shown in Fig. 7.3C. It is called the *electrolytic tilt sensor*. A small slightly curved glass tube is filled with a partly conductive electrolyte. Three electrodes are built into the tube: two at the ends, and the third electrode at the center of the tube. An air bubble resides in the tube and may move along its length as the tube tilts. Electrical resistances between the center electrode and each of the end electrodes depend on the position of the bubble. As the tube shifts away from the balance position, the resistances increase or decrease proportionally. The electrodes are connected into a bridge circuit which is excited with an ac current to avoid damage to the electrolyte and electrodes.

The electrolytic tilt sensors are available[2] for a wide spectrum of angular ranges from $\pm 1^\circ$ to $\pm 80^\circ$. Correspondingly, the shapes of the glass tubes vary from slightly curved to doughnutlike.

A more advanced inclination sensor employs an array of photodetectors [1]. The detector is useful in civil and mechanical engineering for the shape measurements of complex objects with high resolution. Examples include the measurement of ground

[2] The Fredericks Company, Huntingdon Valley, PA.

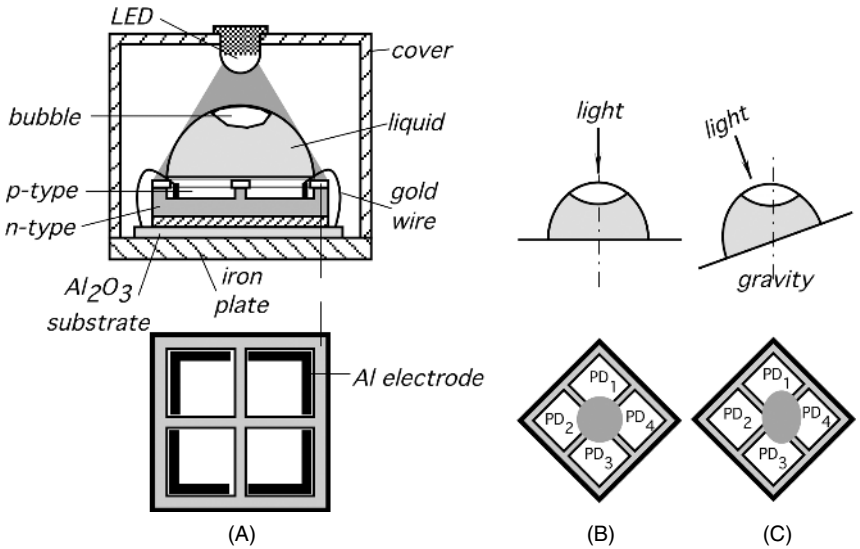


Fig. 7.4. Optoelectronic inclination sensor: (A) design; (B) a shadow at a horizontal position; (C) a shadow at the inclined position.

and road shapes and the flatness of an iron plate, which cannot be done by conventional methods. The sensor (Fig. 7.4A) consists of a light-emitting diode (LED) and a hemispherical spirit level mounted on a p-n-junction photodiode array. A shadow of the bubble in the liquid is projected onto the surface of the photodiode array. When the sensor is kept horizontal, the shadow on the sensor is circular, as shown in Fig. 7.4B, and the area of the shadow on each photodiode of the array is the same. However, when the sensor is inclined, the shadow becomes slightly elliptic, as shown in Fig. 7.4C, implying that the output currents from the diodes are no longer equal. In a practical sensor, the diameter of the LED is 10 mm and the distance between the LED and the level is 50 mm, and the diameters of the hemispherical glass and the bubble are 17 and 9 mm, respectively. The outputs of the diodes are converted into digital form and calibrated at various tilt angles. The calibration data are compiled into look-up tables which are processed by a computing device. By positioning the sensor at the cross point of the lines drawn longitudinally and latitudinally at an interval on the slanting surface of an object, x and y components of the tilt angle can be obtained and the shape of the object is reconstructed by a computer.

7.3 Capacitive Sensors

The capacitive displacement sensors have very broad applications, they are employed directly to gauge displacement and position and also as building blocks in other sensors where displacements are produced by force, pressure, temperature, and so forth. The ability of capacitive detectors to sense virtually all materials makes them an attractive choice for many applications. Equation (3.20) of Chapter 3 states that the capacitance

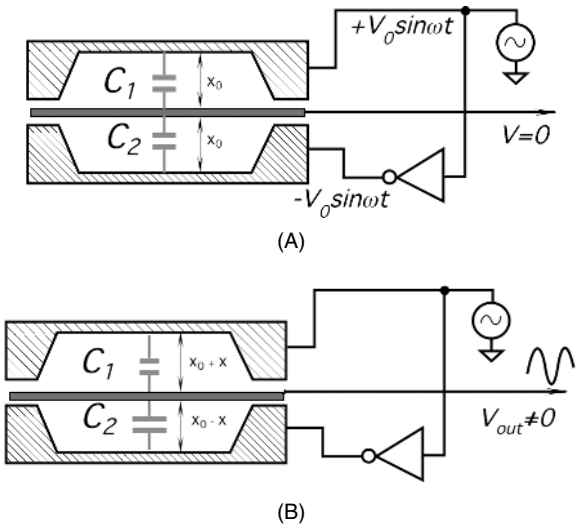


Fig. 7.5. Operating principle of a flat plate capacitive sensor A-balanced position; B-disbalanced position

of a flat capacitor is inversely proportional to the distance between the plates. The operating principle of a capacitive gauge, proximity, and position sensors is based on either changing the geometry (i.e., a distance between the capacitor plates) or capacitance variations in the presence of conductive or dielectric materials. When the capacitance changes, it can be converted into a variable electrical signal. As with many sensors, a capacitive sensor can be either monopolar (using just one capacitor) or differential (using two capacitors), or a capacitive bridge can be employed (using four capacitors). When two or four capacitors are used, one or two capacitors may be either fixed or variable with the opposite phase.

As an introductory example consider three equally spaced plates, each of area A (Fig. 7.5A). The plates form two capacitors C_1 and C_2. The upper and lower plates are fed with the out-of-phase sine-wave signals; that is, the signal phases are shifted by 180°. Both capacitors nearly equal one another and thus the central plate has almost no voltage because the currents through C_1 and C_2 cancel each other. Now, let us assume that the central plate moves downward by a distance x (Fig. 7.5B). This results in changes in the respective capacitance values:

$$C_1 = \frac{\varepsilon A}{x_0 + x} \quad \text{and} \quad C_2 = \frac{\varepsilon A}{x_0 - x}, \tag{7.3}$$

and the central plate signal increases in proportion to the displacement and the phase of that signal is an indication of the central plate direction—up or down. The amplitude of the output signals is

$$V_{\text{out}} = V_0 \left(-\frac{x}{x_0 + x} + \frac{\Delta C}{C} \right). \tag{7.4}$$

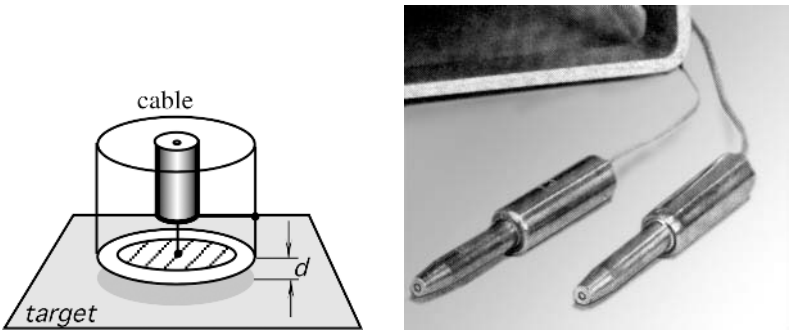


Fig. 7.6. A capacitive probe with a guard ring: (A) cross-sectional view; (B) outside view. (Courtesy of ADE Technologies, Inc., Newton, MA.)

As long as $x \ll x_0$, the output voltage may be considered a linear function of displacement. The second summand represents an initial capacitance mismatch and is the prime cause for the output offset. The offset is also caused by the fringing effects at the peripheral portions of the plates and by the so-called electrostatic force. The force is a result of the charge attraction and repulsion applied to the plates of the sensor, and the plates behave like springs. The instantaneous value of the force is

$$F = -\frac{1}{2}\frac{CV^2}{x_0 + x}. \tag{7.5}$$

In many practical applications, when measuring distances to an electrically conductive object, the object's surface itself may serve as the capacitor's plate. The design of a monopolar capacitive sensor is shown in Fig. 7.6, where one plate of a capacitor is connected to the central conductor of a coaxial cable and the other plate is formed by a target (object). Note that the probe plate is surrounded by a grounded guard to minimize a fringing effect and improve linearity. A typical capacitive probe operates at frequencies in the 3-MHz range and can detect very fast-moving targets, as a frequency response of a probe with a built-in electronic interface is in the range of 40 kHz. A capacitive proximity sensor can be highly efficient when used with the electrically conductive objects. The sensor measures a capacitance between the electrode and the object. Nevertheless, even for the nonconductive objects, these sensors can be employed quite efficiently, although with a lower accuracy. Any object, conductive or nonconductive, that is brought in the vicinity of the electrode, has its own dielectric properties that will alter the capacitance between the electrode and the sensor housing and, in turn, will produce the measurable response.

To improve sensitivity and reduce fringing effects, the monopolar capacitive sensor may be supplied with a driven shield. Such a shield is positioned around the nonoperating sides of the electrode and is fed with the voltage equal to that of the electrode. Because the shield and the electrode voltages are inphase and have the same magnitude, no electric field exists between the two and all components posi-

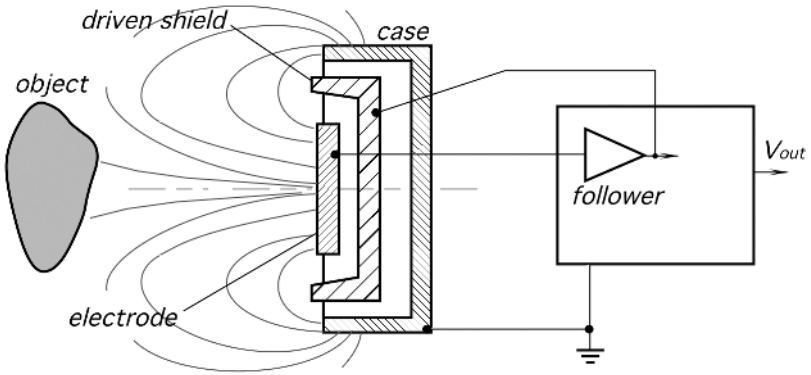


Fig. 7.7. Driven shield around the electrode in a capacitive proximity sensor.

tioned behind the shield have no effect on the operation. The driven-shield technique is illustrated in Fig. 7.7.

Currently, the capacitive bridge became increasingly popular in the design of displacement sensors [2]. A linear bridge capacitive position sensor [3] is shown in Fig. 7.8A. The sensor comprises two planar electrode sets that are parallel and adjacent to each other with a constant separation distance, d. The increase the capacitance, the spacing between the plate sets is relatively small. A stationary electrode set contains four rectangular elements, whereas a moving electrode set contains two rectangular elements. All six elements are of about the same size (a side dimension is b). The size of each plate can be as large as is mechanically practical when a large range of linearity is desired. The four electrodes of the stationary set are cross-connected electrically, thus forming a bridge-type capacitance network.

A bridge excitation source provides a sinusoidal voltage (5–50 kHz) and the voltage difference between the pair of moving plates is sensed by the differential amplifier whose output is connected to the input of a synchronous detector. The capacitance of two parallel plates, of fixed separation distance, is proportional to the area of either plate which directly faces the corresponding area of the other plate. Figure 7.8B shows the equivalent circuit of the sensor which has a configuration of a capacitive bridge. A value of capacitor C_1 is

$$C_1 = \frac{\varepsilon_0 b}{d}\left(\frac{L}{2} + x\right). \tag{7.6}$$

The other capacitances are derived for the identical equations. Note that the opposite capacitors are nearly equal: $C_1 = C_3$ and $C_2 = C_4$. A mutual shift of the plates with respect to a fully symmetrical position results in the bridge disbalance and the phase-sensitive output of the differential amplifier. An advantage of the capacitive bridge circuit is the same as of any bridge circuit: linearity and noise immunity. In addition to the flat electrodes as described earlier, the same method can be applied any symmetrical arrangement of the sensor, (e.g., to detect a rotary motion).

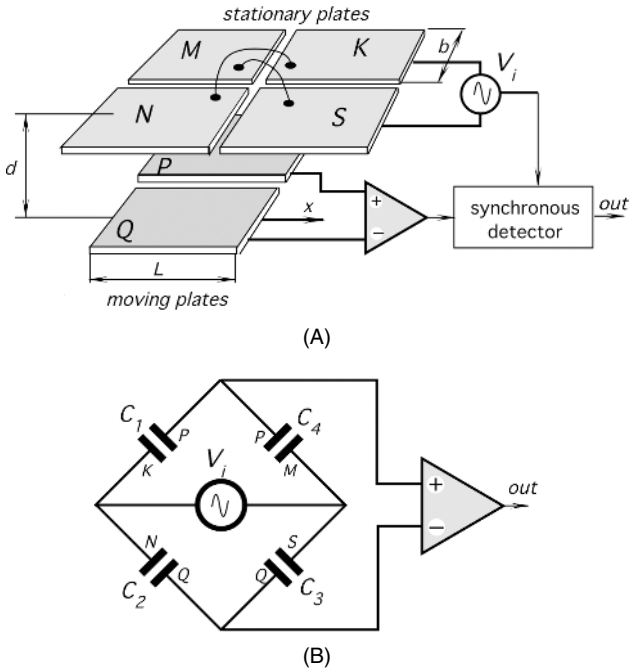


Fig. 7.8. Parallel-plate capacitive bridge sensor: (A) plate arrangement, (B) equivalent circuit diagram.

7.4 Inductive and Magnetic Sensors

One of many advantages of using magnetic field for sensing position and distance is that any nonmagnetic material can be penetrated by the field with no loss of position accuracy. Stainless steel, aluminum, brass, copper, plastics, masonry, and woods can be penetrated, meaning that the accurate position with respect to the probe at the opposite side of a wall can be determined almost instantly. Another advantage is the magnetic sensors can work in severe environments and corrosive situations because the probes and targets can be coated with inert materials that will not adversely affect the magnetic fields.

7.4.1 LVDT and RVDT

Position and displacement may be sensed by methods of electromagnetic induction. A magnetic flux coupling between two coils may be altered by the movement of an object and subsequently converted into voltage. Variable-inductance sensors that use a nonmagnetized ferromagnetic medium to alter the reluctance (magnetic resistance) of the flux path are known as variable-reluctance transducers [4]. The basic arrangement of a multi-induction transducer contains two coils: primary and secondary. The primary carries ac excitation (V_{ref}) that induces a steady ac voltage in the secondary coil (Fig. 7.9). The induced amplitude depends on flux coupling between the coils.

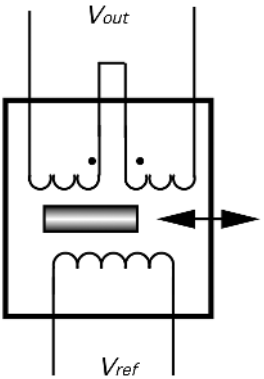


Fig. 7.9. Circuit diagram of the LVDT sensor.

There are two techniques for changing the coupling. One is the movement of an object made of ferromagnetic material within the flux path. This changes the reluctance of the path, which, in turn, alters the coupling between the coils. This is the basis for the operation of a LVDT (linear variable differential transformer), a RVDT (rotary variable differential transformer), and the mutual inductance proximity sensors. The other method is to physically move one coil with respect to another.

The LVDT is a transformer with a mechanically actuated core. The primary coil is driven by a sine wave (excitation signal) having a stabilized amplitude. The sine wave eliminates error-related harmonics in the transformer [5]. An ac signal is induced in the secondary coils. A core made of a ferromagnetic material is inserted coaxially into the cylindrical opening without physically touching the coils. The two secondaries are connected in the opposed phase. When the core is positioned in the magnetic center of the transformer, the secondary output signals cancel and there is no output voltage. Moving the core away from the central position unbalances the induced magnetic flux ratio between the secondaries, developing an output. As the core moves, the reluctance of the flux path changes. Hence, the degree of flux coupling depends on the axial position of the core. At a steady state, the amplitude of the induced voltage is proportional, in the linear operating region, to the core displacement. Consequently, voltage may be used as a measure of a displacement. The LVDT provides the direction as well as magnitude of the displacement. The direction is determined by the phase angle between the primary (reference) voltage and the secondary voltage. Excitation voltage is generated by a stable oscillator. To exemplify how the sensor works, Fig. 7.10 shows the LVDT connected to a synchronous detector which rectifies the sine wave and presents it at the output as a dc signal. The synchronous detector is composed of an analog multiplexer (MUX) and a zero-crossing detector which converts the sine wave into the square pulses compatible with the control input of the multiplexer. A phase of the zero-crossing detector should be trimmed for the zero output at the central position of the core. The output amplifier can be trimmed to a desirable gain to make the signal compatible with the next stages. The synchronized clock to the multiplexer means that the information presented to the RC filter at the input of the amplifier is amplitude and phase sensitive. The output voltage represents how far the core is from the center and on which side.

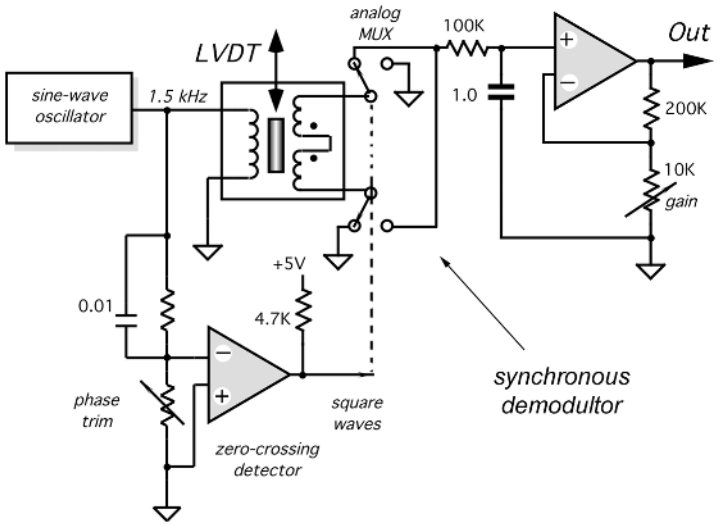


Fig. 7.10. A simplified circuit diagram of an interface for an LVDT sensor.

For the LVDT to measure transient motions accurately, the frequency of the oscillator must be at least 10 times higher than the highest significant frequency of the movement. For the slow-changing process, stable oscillator may be replaced by coupling to a power line frequency of 60 or 50 Hz.

Advantages of the LVDT and RVDT are the following: (1) The sensor is a noncontact device with no or very little friction resistance with small resistive forces; (2) hystereses (magnetic and mechanical) are negligible; (3) output impedance is very low; (4) there is low susceptibility to noise and interferences; (5) its construction is solid and robust, (6) infinitesimal resolution is possible.

One useful application for the LVDT sensor is in the so-called *gauge heads*, which are used in tool inspection and gauging equipment. In that case, the inner core of the LVDT is spring loaded to return the measuring head to a preset reference position.

The RVDT operates on the same principle as LVDT, except that a rotary ferromagnetic core is used. The prime use for the RVDT is the measurement of angular displacement. The linear range of measurement is about $\pm 40°$, with a nonlinearity error of about 1%.

7.4.2 Eddy Current Sensors

To sense the proximity of nonmagnetic but conductive materials, the effect of *eddy currents* is used in a dual-coil sensor (Fig. 7.11A). One coil is used as a reference, and the other is for the sensing of the magnetic currents induced in the conductive object. Eddy (circular) currents produce a magnetic field which opposes that of the sensing coil, thus resulting in a disbalance with respect to the reference coil. The closer the

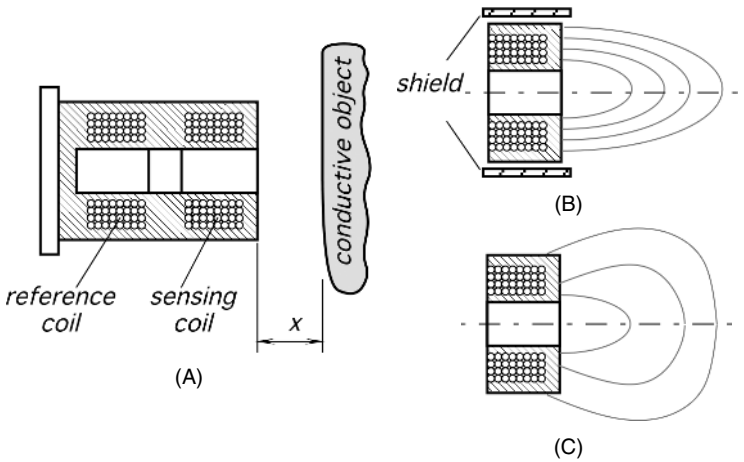


Fig. 7.11. (A) Electromagnetic proximity sensor; (B) sensor with the shielded front end; (C) unshielded sensor.

object to the coil, the larger the change in the magnetic impedance. The depth of the object where eddy currents are produced is defined by

$$\delta = \frac{1}{\sqrt{\pi f \mu \sigma}}, \tag{7.7}$$

where f is the frequency and σ is the target conductivity. Naturally, for effective operation, the object thickness should be larger than the depth. Hence, eddy detectors should not be used for detecting metallized film or foil objects. Generally, the relationship between the coil impedance and distance to the object x is nonlinear and temperature dependent. The operating frequency of the eddy current sensors range from 50 kHz to 10 MHz.

Figures 7.11B and 7.11C show two configurations of the eddy sensors: with the shield and without one. The shielded sensor has a metal guard around the ferrite core and the coil assembly. It focuses the electromagnetic field to the front of the sensor. This allows the sensor to be imbedded into a metal structure without influencing the detection range. The unshielded sensor can sense at its sides as well as from the front. As a result, the detecting range of an unshielded sensor is usually somewhat greater than that of the shielded sensor of the same diameter. To operate properly, the unshielded sensors require nonmetallic surrounding objects.

In addition to position detection, eddy sensors can be used to determine material thickness, nonconductive coating thickness, conductivity and plating measurements, and cracks in the material. Crack detection and surface flaws become the most popular applications for the sensors. Depending on the applications, eddy probes may be of many coil configurations: Some are very small in diameter (2–3 mm) and others are quite large (25 mm). Some companies even make custom-designed probes to meet

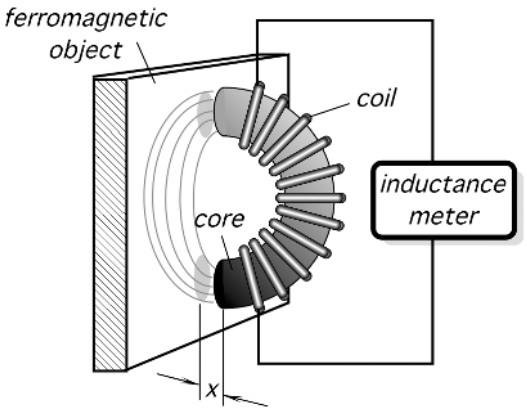


Fig. 7.12. A transverse inductive proximity sensor.

unique requirements of the customers (Staveley Instruments, Inc., Kennewick, WA). One important advantage of the eddy current sensors is that they do not need magnetic material for the operation, thus they can be quite effective at high temperatures (well exceeding the Curie temperature of a magnetic material) and for measuring the distance to or level of conductive liquids, including molten metals. Another advantage of the detectors is that they are not mechanically coupled to the object and, thus, the loading effect is very low.

7.4.3 Transverse Inductive Sensor

Another position-sensing device is called a *transverse inductive proximity sensor*. It is useful for sensing relatively small displacements of ferromagnetic materials. As the name implies, the sensor measures the distance to an object which alters the magnetic field in the coil. The coil inductance is measured by an external electronic circuit (Fig. 7.12). A self-induction principle is the foundation for the operation of such a transducer. When the proximity sensor moves into the vicinity of a ferromagnetic object, its magnetic field changes, thus altering the inductance of the coil. The advantage of the sensor is that it is a noncontact device whose interaction with the object is only through the magnetic field. An obvious limitation is that it is useful only for the ferromagnetic objects at relatively short distances.

A modified version of the transverse transducer is shown in Fig. 7.13A. To overcome the limitation for measuring only ferrous materials, a ferromagnetic disk is attached to a displacing object while the coil is in a stationary position. Alternatively, the coil may be attached to the object and the core is stationary. This proximity sensor is useful for measuring small displacements only, as its linearity is poor in comparison with the LVDT. However, it is quite useful as a proximity detector for the indication of the close proximity to an object which is made of any solid material. The magnitude of the output signal as function of distance to the disk is shown in Fig. 7.13B.

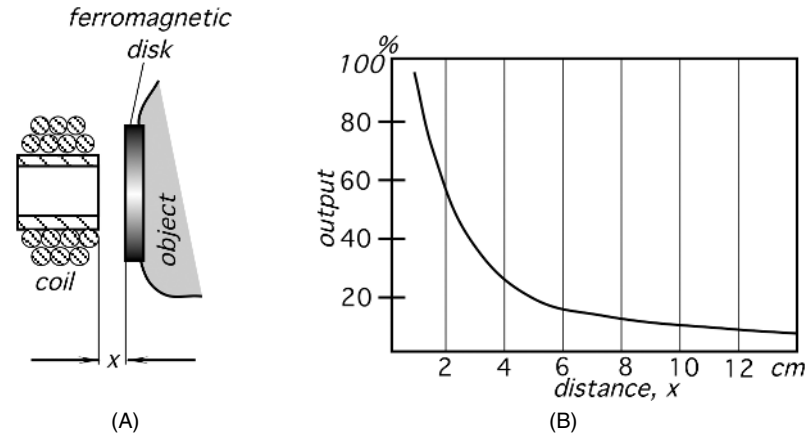


Fig. 7.13. Transverse sensor with an auxiliary ferromagnetic disk (A) and the output signal as function of distance (B).

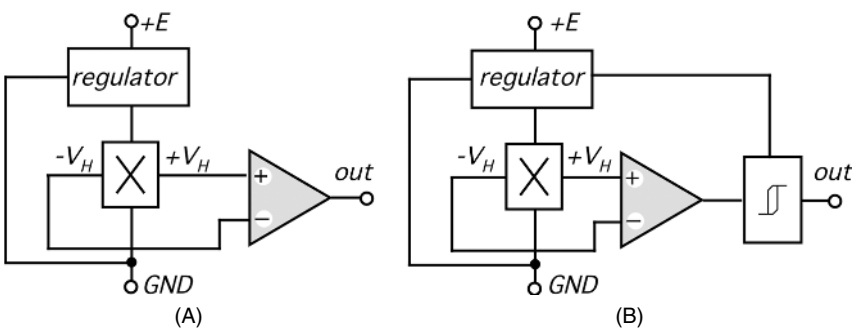


Fig. 7.14. Circuit diagrams of a linear (A) and a threshold (B) Hall effect sensor.

7.4.4 Hall Effect Sensors

During recent years, Hall effect sensors became increasingly popular.[3] There are two types of Hall sensors: linear and threshold (Fig. 7.14). A linear sensor usually incorporates an amplifier for the easier interface with the peripheral circuits. In comparison with a basic sensor (Fig. 3.30 of Chapter 3), they operate over a broader voltage range and are more stable in a noisy environment. These sensors are not quite linear (Fig.7.15A) with respect to magnetic field density and, therefore, the precision measurements require a calibration. In addition to the amplifier, the threshold-type sensor contains a Schmitt trigger detector with a built-in hysteresis. The output signal as a function of a magnetic field density is shown in Fig. 7.15B. The signal is a two-level one and has clearly pronounced hysteresis with respect to the magnetic field. When the applied magnetic flux density exceeds a certain threshold, the trigger provides a

[3] See Section 3.8 of Chapter 3 for the operating principle.

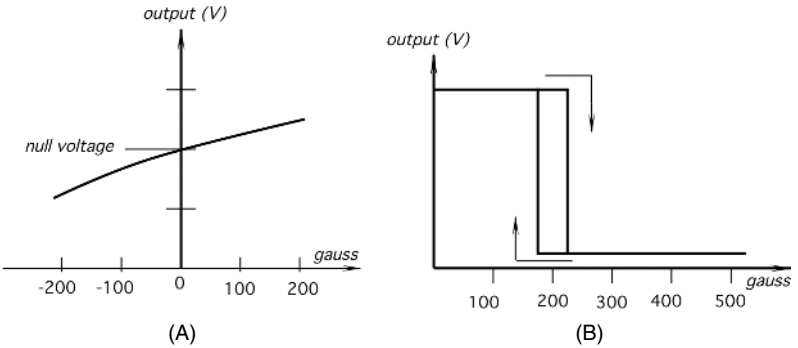


Fig. 7.15. Transfer functions of a linear (A) and a threshold (B) Hall effect sensor.

clean transient from the OFF to the ON position. The hysteresis eliminates spurious oscillations by introducing a dead-band zone, in which the action is disabled after the threshold value has passed. The Hall sensors are usually fabricated as monolithic silicon chips and encapsulated into small epoxy or ceramic packages.

For the position and displacement measurements, the Hall effect sensors must be provided with a magnetic field source and an interface electronic circuit. The magnetic field has two important characteristics for this application: a flux density and a polarity (or orientation). It should be noted that for better responsivity, magnetic field lines must be normal (perpendicular) to the flat face of the sensor and must be at the correct polarity. In the Sprague® threshold sensors, the south magnetic pole will cause switching action and the north pole will have no effect.

Before designing a position detector with a Hall sensor, an overall analysis should be performed in approximately the following manner. First, the field strength of the magnet should be investigated. The strength will be the greatest at the pole face and will decrease with increasing distance from the magnet. The field may be measured by a gaussmeter or a calibrated Hall sensor. For the threshold-type Hall sensor, the longest distance at which the sensor's output goes from ON (high) to OFF (low) is called a *release point*. It can be used to determine the critical distance where the sensor is useful. The magnetic field strength is not linear with distance and depends greatly on the magnet shape, the magnetic circuit, and the path traveled by the magnet. The Hall conductive strip is situated at some depth within the sensor's housing. This determines the minimum operating distance. A magnet must operate reliably with the total effective air gap in the working environment. It must fit the available space and must be mountable, affordable, and available.[4]

The Hall sensors can be used for interrupter switching with a moving object. In this mode, the activating magnet and the Hall sensor are mounted on a single rugged assembly with a small air gap between them (Fig. 7.16). Thus, the sensor is held in the ON position by the activating magnet. If a ferromagnetic plate, or vane, is placed between the magnet and the Hall sensor, the vane forms a magnetic shunt that distorts the magnetic flux away from the sensor. This causes the sensor to flip to

[4] For more information on permanent magnets, see Section 3.4 of Chapter 3.

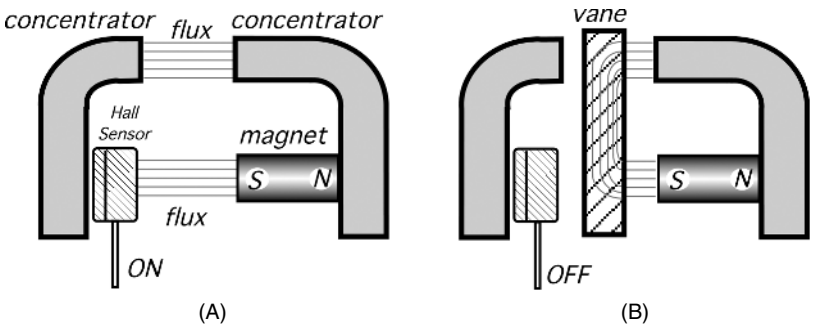


Fig. 7.16. The Hall effect sensor in the interrupter switching mode: (A) the magnetic flux turns the sensor on; (B) the magnetic flux is shunted by a vane. (After Ref. [6].)

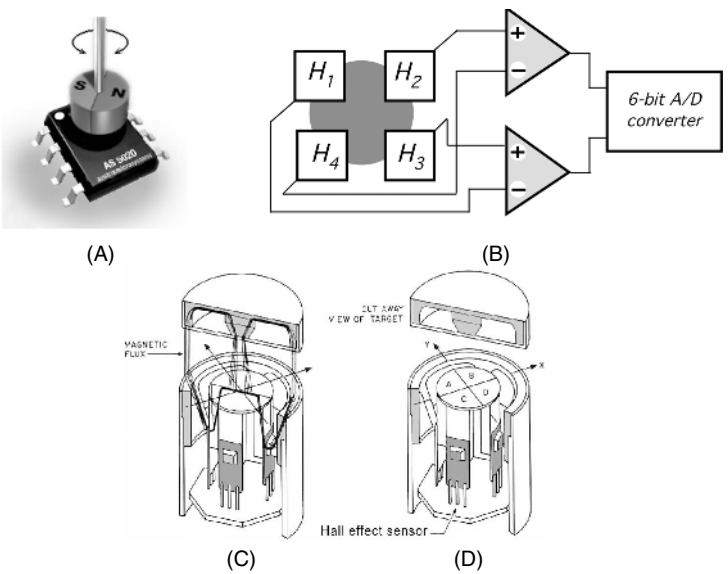


Fig. 7.17. Angular Hall sensor bridge (A) and the internal sensor interface (B) (Courtesy of Austria Micro Systems). A cut-away view (C) of the sensor with the target and the probe shows the magnetic flux paths. A cut-away view (D) shows four Hall effect sensors with four flux return paths.

the OFF position. The Hall sensor and the magnet could be molded into a common housing, thus eliminating the alignment problem. The ferrous vanes which interrupt the magnetic flux could have linear or rotating motion. An example of such a device is an automobile distributor.

Like many other sensors, four Hall sensors can be connected into a bridge circuit to detect linear or circular motion. Figures 7.17A and 7.17B illustrate this concept where the sensor is fabricated using MEMS technology on a single chip and packaged in a

SOIC-8 plastic housing. A circular magnet is positioned above the chip and its angle of rotation and direction is sensed and converted into a digital code. The properties of an analog-to-digital converter determine the speed response that allows the magnet to rotate with a rate of up to 30,000 rpm. Such a sensor permits a friction-free precision linear and angular sensing of position, precision angular encoding, and even making a programmable rotary switch. Because of a bridge connection of the individual sensor, the circuit is highly tolerant of the magnet's misalignment and external interferences, including the magnetic fields.

The design of a three-dimensional (3-D) coordinate Hall effect sensor works by electronically measuring and comparing the magnetic flux from a movable target through four geometrically equal magnetic paths arranged symmetrically around the axis of the probe (Figs. 7.17C and 7.17D). It is a magnetic equivalent of a Wheatstone bridge. The target's symmetrical magnetic field, generated by a permanent magnet, travels from the central pole through the air to the outer rim, when it is not in the vicinity of the probe. Because the flux from the target will take the path of least resistance (reluctance), the flux will go through the probe when the target is sufficiently close to it. The probe has a central pole face divided into four equal sections. The values of flux in the A, B, C, and D paths are measured by the respective Hall effect sensors. There are two ways to fabricate a target. One is active and the other is passive. An active target uses a permanent magnet to generate a magnetic field, which is sensed by the probe when it is within the operating range. A passive target does not generate a magnetic field; instead, the field is generated by the probe and returned by the target. An example of the application is the unmanned vehicle guidance system that leads a vehicle over a roadbed with passive metal strip targets buried just under the road surface. The probe is attached to the vehicle. The targets will give position, speed, and direction as the probe passes over it. A probe and target can be separated by several inches.

As shown in Figs. 7.17A and 7.17B, a rotary motion can be digitally encoded with high precision. To take advantage of this feature, a linear distance sensor can be built with a converter of a linear into a rotary motion as shown in Fig. 7.18. Such sensors are produced, for example, by SpaceAge Control, Inc. (www.spaceagecontrol.com).

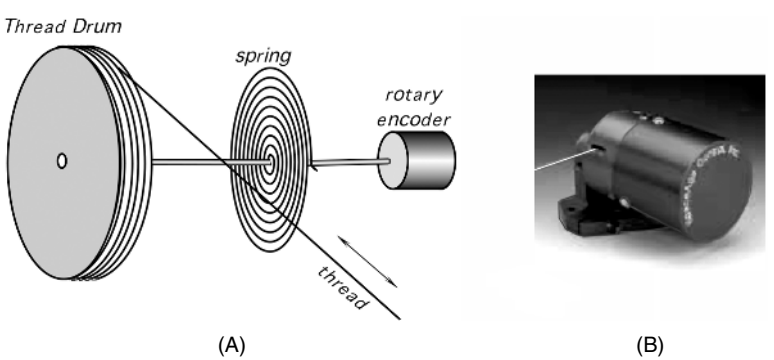


Fig. 7.18. Conversion of a linear displacement (length of a thread or cable) into a rotary motion (A) and cable position sensor (B). (Courtesy of Space Age Control, Inc.)

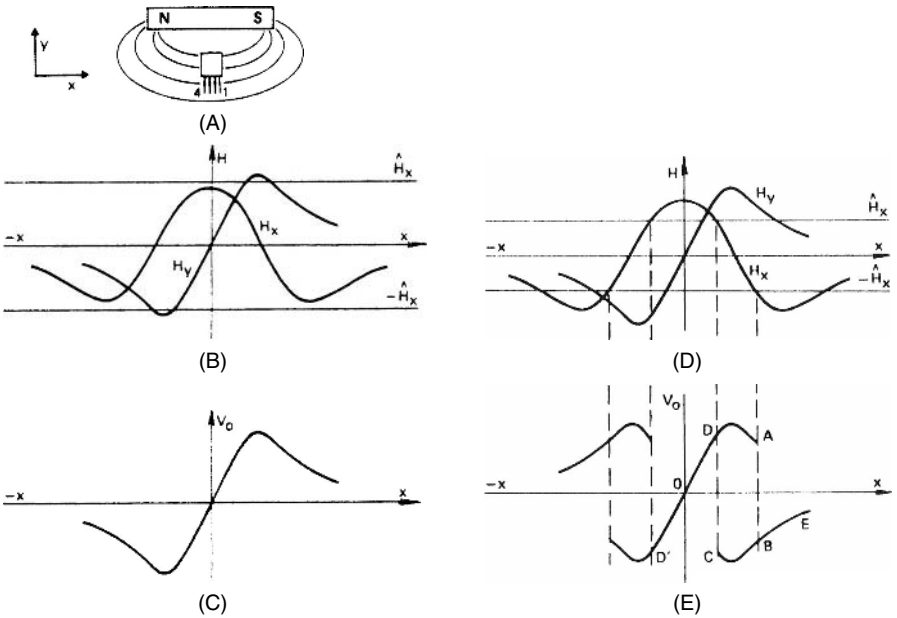


Fig. 7.19. Magnetoresistive sensor output in the field of a permanent magnet as a function of its displacement x parallel to the magnetic axis (A–C). The magnet provides both the axillary and transverse fields. Reversal of the sensor relative to the magnet will reverse the characteristic. (D and E) Sensor output with a too strong magnetic field.

7.4.5 Magnetoresistive Sensors[5]

These sensors are similar in application to the Hall effect sensors. For functioning, they require an external magnetic field. Hence, whenever the magnetoresistive sensor is used as a proximity, position, or rotation detector, it must be combined with a source of a magnetic field. Usually, the field is originated in a permanent magnet which is attached to the sensor. Figure 7.19 shows a simple arrangement for using a sensor–permanent-magnet combination to measure linear displacement. It reveals some of the problems likely to be encountered if proper account is not taken of the effects described in this subsection. When the sensor is placed in the magnetic field, it is exposed to the fields in both the x and y directions. If the magnet is oriented with its axis parallel to the sensor strips (i.e., in the x direction) as shown in Fig. 7.19A, $\mathbf{H}_x$ then provides the auxiliary field, and the variation in $\mathbf{H}_y$ can be used as a measure of x displacement. Figure 7.19B shows how both $\mathbf{H}_x$ and $\mathbf{H}_y$ vary with x, and Fig. 7.19C shows the corresponding output signal. In this example, $\mathbf{H}_x$ never exceeds $\pm\mathbf{H}_x$ (the field that can cause flipping of the sensor), and the sensor characteristics remain stable and well behaved throughout the measuring range. However, if the magnet is

[5] Information on the KZM10 and KM110 sensors is courtesy of Philips Semiconductors BV (Eindhoven, The Netherlands).

too powerful or the sensor passes too close to the magnet, the output signal will be drastically different.

Suppose the sensor is initially on the transverse axis of the magnet ($x = 0$). $\mathbf{H}_y$ will be zero and $\mathbf{H}_x$ will be at its maximum value ($> \mathbf{H}_x$). Thus, the sensor will be oriented in the $+x$ direction and the output voltage will vary as in Fig. 7.19E. With the sensor's movement in the $+x$ direction, $\mathbf{H}_y$ and V_0 increase, and $\mathbf{H}_x$ falls to zero and then increases negatively until $\mathbf{H}_y$ exceeds $-\mathbf{H}_x$. At this point, the sensor characteristic flips and the output voltage reverses, moving from A to B in Fig. 7.19E. A further increase in x causes the sensor voltage to move along BE. If the sensor is moved in the opposite direction, however, $\mathbf{H}_x$ increases until it exceeds $+\mathbf{H}_x$ and V_0 moves from B to C. At this point, the sensor characteristic again flips and V_0 moves from C to D. Then, under these conditions, the sensor characteristic will trace the hysteresis loop ABCD and a similar loop in the $-x$ direction. Figure 7.19E is an idealized case, because the reversals are never as abrupt as shown.

Figure 7.20A shows how KMZ10B and KM110B magnetoresistive sensors may be used to make position measurements of a metal object. The sensor is located between the plate and a permanent magnet, which is oriented with its magnetic axis

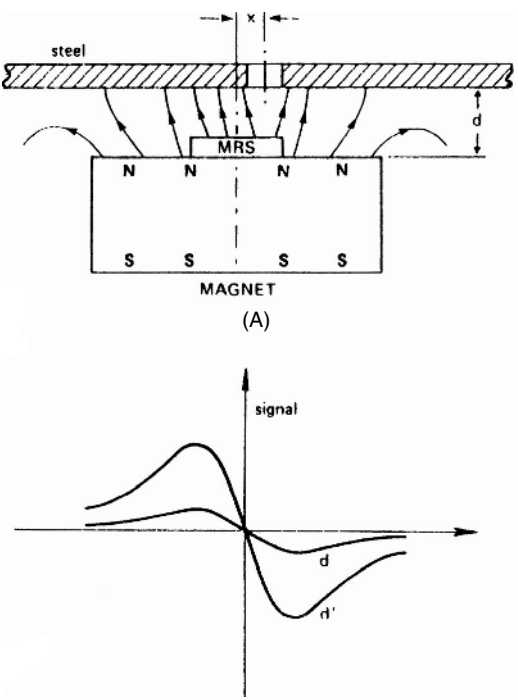


Fig. 7.20. One point measurement with the KMZ10. (A) The sensor is located between the permanent magnet and the metal plate; (B) Output signals for two distances between the magnet and the plate.

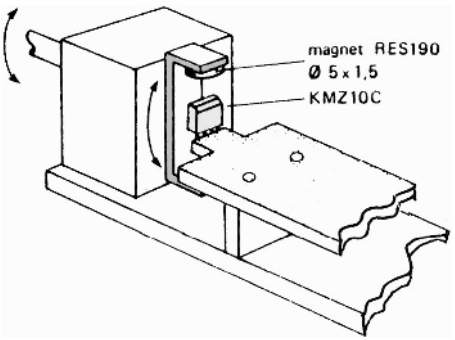


Fig. 7.21. Angular measurement with the KMZ10 sensor.

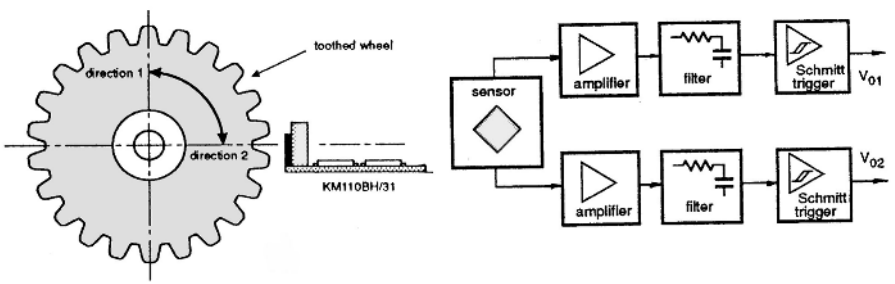


Fig. 7.22. (A) Optimum operating position of a magnetoresistive module. Note a permanent magnet positioned behind the sensor. (B) Block diagram of the module circuit.

normal to the axis of the metal plate. A discontinuity in the plate's structure, such as a hole or a region of nonmagnetic material, will disturb the magnetic field and produce a variation in the output signal from the sensor. Figure 7.20B shows the output signal for two values of spacing d. At the point where the hole and the sensor are precisely aligned, the output is zero regardless of the distance d or surrounding temperature.

Figure 7.21 shows another setup which is useful for measuring angular displacement. The sensor itself is located in the magnetic field produced by two RES190 permanent magnets fixed to a rotable frame. The output of the sensor will then be a measure of the rotation of the frame.

Figure 7.22A depicts the use of a single KM110 sensor for detecting rotation and direction of a toothed wheel. The method of direction detection is based on a separate signal processing for the sensor's two half-bridge outputs.

The sensor operates like a magnetic Wheatstone bridge measuring nonsymmetrical magnetic conditions such as when the teeth or pins move in front of the sensor. The mounting of the sensor and the magnet is critical, so the angle between the sensor's symmetry axis and that of the toothed wheel must be kept near zero. Further, both axes (the sensor's and the wheel's) must coincide. The circuit (Fig. 7.22B) connects both bridge outputs to the corresponding amplifiers and, subsequently, to the low-pass filters and Schmitt triggers to form the rectangular output signals. A phase difference between both outputs (Figs. 7.23A and 7.23B) is an indication of a rotation direction.

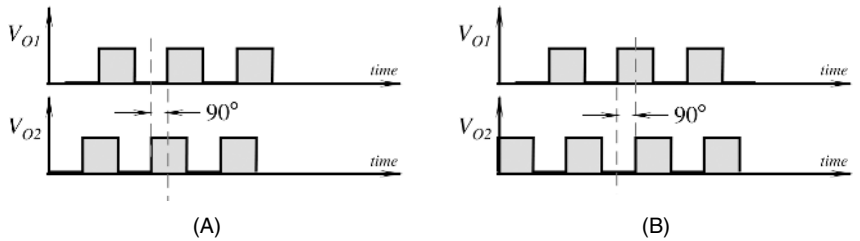


Fig. 7.23. Output signal from the amplifiers for direction 1 (A) and 2 (B).

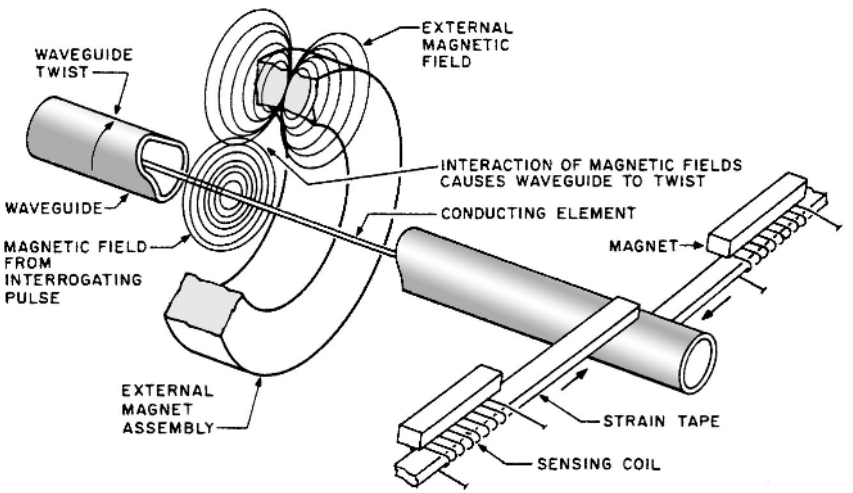


Fig. 7.24. A magnetostrictive detector uses ultrasonic waves to detect position of a permanent magnet.

7.4.6 Magnetostrictive Detector

A transducer which can measure displacement with high resolution across long distances can be built by using magnetostrictive and ultrasonic technologies [8]. The transducer is comprised of two major parts: a long waveguide (up to 7 m long) and a permanent ring magnet (Fig. 7.24). The magnet can move freely along the waveguide without touching it. A position of that magnet is the stimulus which is converted by the sensor into an electrical output signal. A waveguide contains a conductor which, upon applying an electrical pulse, sets up a magnetic field over its entire length. Another magnetic field produced by the permanent magnet exists only in its vicinity. Thus, two magnetic fields may be setup at the point where the permanent magnet is located. A superposition of two fields results in the net magnetic field, which can be found from the vector summation. This net field, although helically formed around the waveguide, causes it to experience a minute torsional strain, or twist at the location of the magnet. This twist is known as the Wiedemann effect.

Therefore, electric pulses injected into the waveguide's coaxial conductor produce mechanical twist pulses which propagate along the waveguide with the speed of sound specific for its material. When the pulse arrives at the excitation head of the sensor, the moment of its arrival is precisely measured. One way to detect that pulse is to use a detector that can convert an ultrasonic twitch into electric output. This can be accomplished by piezoelectric sensors or, as it is shown in Fig. 7.24, by the magnetic reluctance sensor. The sensor consists of two tiny coils positioned near two permanent magnets. The coils are physically coupled to the waveguide and can jerk whenever the waveguide experiences the twitch. This sets up short electric pulses across the coils. The time delay of these pulses from the corresponding excitation pulses in the coaxial conductor is the exact measure of the ring magnet position. An appropriate electronic circuit converts the time delay into a digital code representative of a position of the permanent magnet on the waveguide. The advantage of this sensor is in its high linearity (on the order of 0.05% of full scale), good repeatability (on the order of 3 μm), and long-term stability. The sensor can withstand aggressive environments, such as high pressure, high temperature, and strong radiation. Another advantage of this sensor is its low-temperature sensitivity which by careful design can be achieved on the order of 20 ppm/°C.

Applications of this sensor include hydraulic cylinders, injection-molding machines (to measure linear displacement for mold clamp position, injection of molding material, and ejection of the molded part), mining (for detection of rocks movements as small as 25 μm), rolling mills, presses, forges, elevators, and other devices where fine resolution along large dimensions is a requirement.

7.5 Optical Sensors

After mechanical contact and potentiometric sensors, optical sensors are probably the most popular for measuring position and displacement. Their main advantages are simplicity, the absence of the loading effect, and relatively long operating distances. They are insensitive to stray magnetic fields and electrostatic interferences, which makes them quite suitable for many sensitive applications. An optical position sensor usually requires at least three essential components: a light source, a photodetector, and light guidance devices, which may include lenses, mirrors, optical fibers, and so forth. An example of single- and dual-mode fiber-optic proximity sensors are shown in Figs. 4.17 and 4.18 of Chapter 4. Similar arrangements are often implemented without optical fibers when light is guided toward a target by focusing lenses and is diverted back to detectors by the reflectors. Currently, this basic technology has been substantially improved. Some more complex and sophisticated products have evolved. The improvements are aimed to better selectivity, noise immunity, and reliability of the optical sensors.

7.5.1 Optical Bridge

The concept of a bridge circuit, like a classical Wheatstone bridge, is employed in many sensors and the optical sensor is a good example of that. One such use shown in

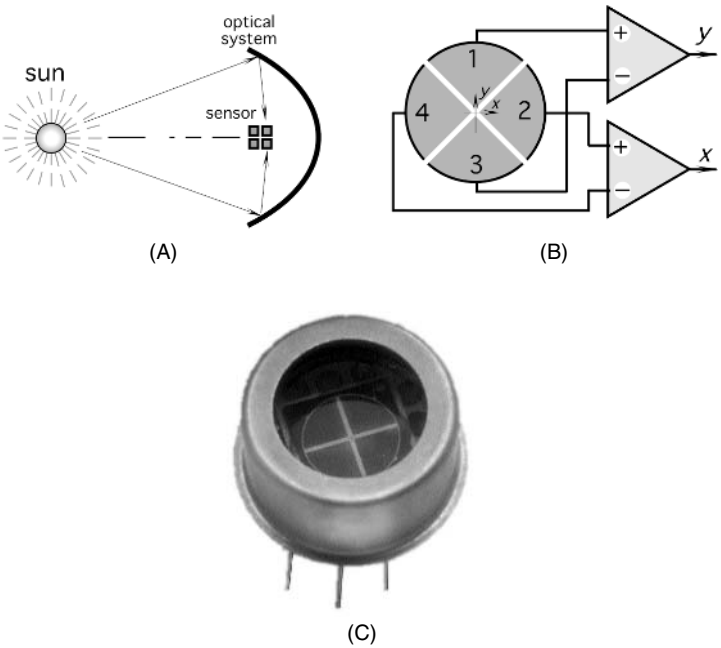


Fig. 7.25. Four-quadrant photodetector: (A) focusing an object on the sensor; (B) connection of the sensing elements to difference amplifiers; (C) sensor in a packaging. (From Advanced Photonix, Inc. Camarillo, CA.)

Fig. 7.25. A four-quadrant photodetector consists of four light detectors connected in a bridgelike circuit. The object must have an optical contrast against the background. Consider a positioning system of a spacecraft (Fig. 7.25A). An image of the Sun or any other sufficiently bright object is focused by an optical system (a telescope) on a four-quadrant photodetector. The opposite parts of the detector are connected to the corresponding inputs of the difference amplifiers (Fig. 7.25B). Each amplifier produces the output signal proportional to a displacement of the image from the optical center of the sensor along a corresponding axis. When the image is perfectly centered, both amplifiers produce zero outputs. This may happen only when the optical axis of the telescope passes through the object.

7.5.2 Proximity Detector with Polarized Light

One method of building a better optoelectronic sensor is to use polarized light. Each light photon has specific magnetic and electric field directions perpendicular to each other and to the direction of propagation (see Fig. 3.48 of Chapter 3). The direction of the electric field is the direction of the light *polarization*. Most of the light sources produce light with randomly polarized photons. To make light polarized, it can be directed through a polarizing filter, (i.e., a special material which transmits light polarized only in one direction and absorbs and reflects photons with wrong polarizations).

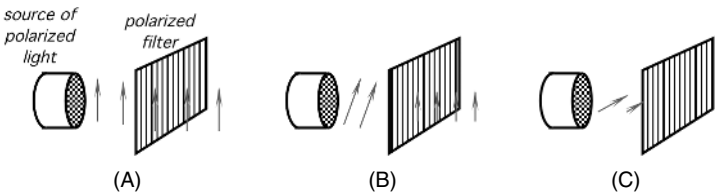


Fig. 7.26. Passing polarized light through a polarizing filter: (A) direction of polarization is the same as of the filter; (B) direction of polarization is rotated with respect to the filter; (C) direction of polarization is perpendicular with respect to the filter.

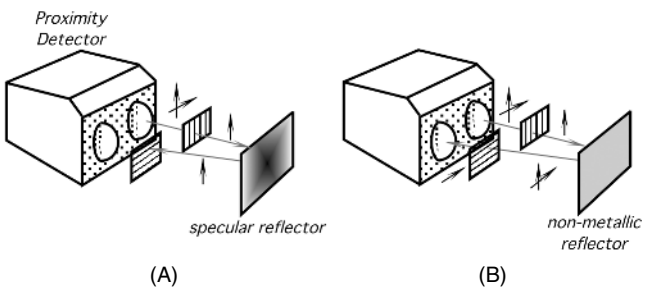


Fig. 7.27. Proximity detector with two polarizing filters positioned at a 90° angle with respect to one another: (A) polarized light returns from the metallic object within the same plane of polarization; (B) nonmetallic object depolarizes light, thus allowing it to pass through the polarizing filter.

However, any direction of polarization can be represented as a geometrical sum of two orthogonal polarizations: One is the same as the filter and the other is nonpassing. Thus, by rotating the polarization of light before the polarizing filter, we may *gradually* change the light intensity at the filter's output (Fig. 7.26).

When polarized light strikes an object, the reflected light may retain its polarization (specular reflection) or the polarization angle may change. The latter is typical for many nonmetallic objects. Thus, to make a sensor nonsensitive to reflective objects (like metal cans, foil wrappers, and the like), it may include two perpendicularly positioned polarizing filters: one at the light source and the other at the detector (Figs. 7.27A and 7.27B). The first filter is positioned at the emitting lens (light source) to polarize the outgoing light. The second filter is at the receiving lens (detector) to allow passage of only those components of light which have a 90° rotation with respect to the outgoing polarization. Whenever light is reflected from a specular reflector, its polarization direction does not change and the receiving filter will not allow the light to pass to a photodetector. However, when light is reflected in a nonspecular manner, its components will contain a sufficient amount of polarization to go through the receiving filter and activate the detector. Therefore, the use of polarizers reduces false-positive detections of nonmetallic objects.

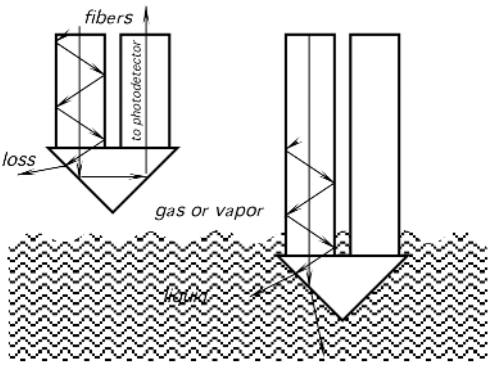


Fig. 7.28. Optical liquid-level detector utilizing a change in the refractive index.

7.5.3 Fiber-Optic Sensors

Fiber-optic sensors can be used quite effectively as proximity and level detectors. One example of the displacement sensor is shown in Fig. 4.18 of Chapter 4, where the intensity of the reflected light is modulated by the distance d to the reflective surface.

A liquid-level detector (see also Section 7.8.3) with two fibers and a prism is shown in Fig. 7.28. It utilizes the difference between refractive indices of air (or gaseous phase of a material) and the measured liquid. When the sensor is above the liquid level, a transmitting fiber (on the left) sends most of its light to the receiving fiber (on the right) due to a total internal reflection in the prism. However, some light rays approaching the prism reflective surface at angles less than the angle of total internal reflection are lost to the surroundings. When the prism reaches the liquid level, the angle of total internal reflection changes because the refractive index of a liquid is higher than that of air. This results in a much greater loss in the light intensity, which can be detected at the other end of the receiving fiber. The light intensity is converted into an electrical signal by any appropriate photodetector. Another version of the sensor is shown in Fig. 7.29, which shows a sensor fabricated by Gems Sensors (Plainville, CT). The fiber is U-shaped, and upon being immersed into liquid, it modulates the intensity of passing light. The detector has two sensitive regions near the bends, where the radius of curvature is the smallest. An entire assembly is packaged into a 5-mm-diameter probe and has a repeatability error of about 0.5 mm. Note that the shape of the sensing element draws liquid droplets away from the sensing regions when the probe is elevated above the liquid level.

7.5.4 Fabry–Perot Sensors

For measuring small displacements with high precision in a harsh environment, the so-called Fabry–Perot optical cavity can be employed. The cavity contains two semireflective mirrors facing each other and separated by distance L (Fig. 7.30A). The cavity is injected with light from a known source (a laser, e.g.) and the photons inside the cavity bounce back and forth between the two mirrors, interfering with each other in the process. In fact, the cavity is a storage tank for light. At some frequencies of photons, light can pass out of the cavity. A Fabry–Perot interferometer is basically

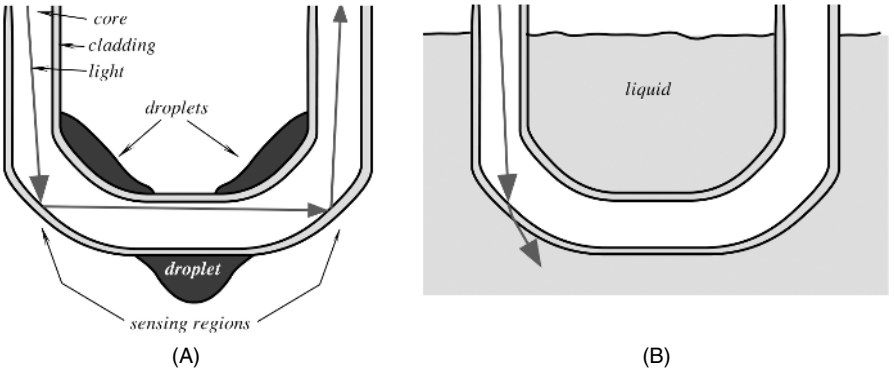


Fig. 7.29. U-shaped fiber-optic liquid-level sensor: (A) When the sensor is above the liquid level, the light at the output is strongest; (B) when the sensitive regions touch liquid, the light propagated through the fiber drops.

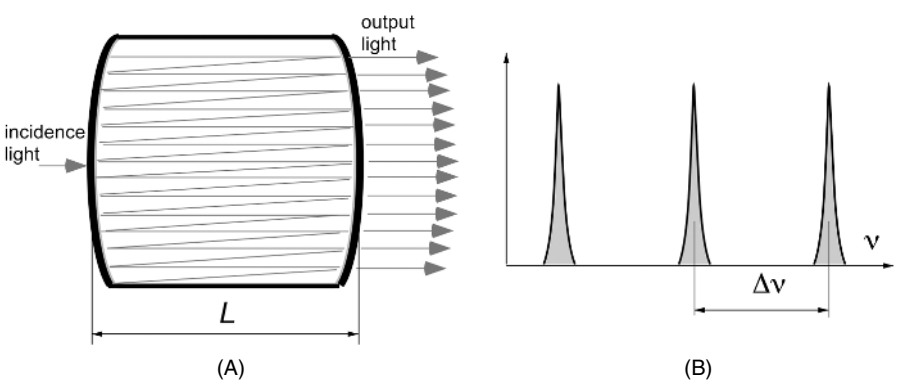


Fig. 7.30. (A) Multiple-ray interference inside a Fabry–Perot cavity; (B) transmitted frequencies of light.

a frequency filter whose transmission frequency is intimately related to the length of the cavity (Fig. 7.30B). As the cavity length changes, the frequencies at which it transmits light change accordingly. If you make one of the mirrors movable, by monitoring the optical transmission frequency, very small changes in the cavity length can be resolved. The narrow bands of transmitted light are separated by frequencies that are inversely proportional to the cavity length:

$$\Delta\nu = \frac{c}{2L}, \tag{7.8}$$

where c is the speed of light. For practical cavities with a mirror separation on the order of 1 μm, typical values of $\Delta\nu$ are between 500 MHz and 1 GHz. Thus, by detecting the frequency shift of the transmitted light with respect to a reference light source, changes in the cavity dimensions can be measured with the accuracy comparable

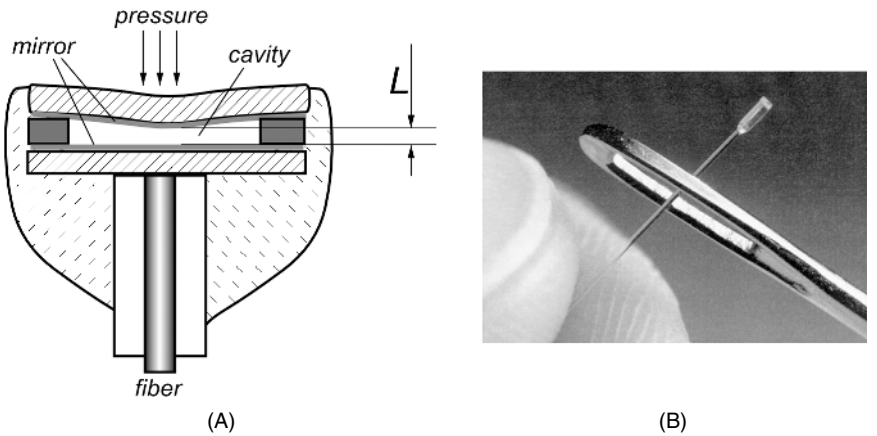


Fig. 7.31. Construction of a Fabry–Perot pressure sensor (A) and view of FISO FOP-M pressure sensor (B).

with the wavelength of light. Whatever may cause changes in the cavity dimensions (mirror movement) may be the subject of measurements. These include strain, force, pressure, and temperature.

Fabry–Perot cavity-based sensors have been widely used for their versatility; for example, they have been used to sense both pressure and temperature [7–10]. This kind of sensor detects changes in optical path length induced by either a change in the refractive index or a change in physical length of the cavity. Micromachining techniques make Fabry–Perot sensors more attractive by reducing the size and the cost of the sensing element. Another advantage of the miniature Fabry–Perot sensor is that low-coherence light sources, such as light-emitting diodes (LEDs) or even light bulbs, can be used to generate the interferometric signal.

A pressure sensor with a Fabry–Perot cavity is shown in Fig. 7.31A. Pressure is applied to the upper membrane. Under pressure, the diaphragm deflects inwardly, thus reducing the cavity dimension L. The cavity is monolithically built by micromachined technology and the mirrors can be either the dielectric layers or metal layers deposited or evaporated during the manufacturing process. The thickness of each layer must be tightly controlled to achieve the target performance of a sensor. An ultraminiature pressure sensor produced by FISO Technologies (www.fiso.com) is shown in Fig. 7.31B. The sensor has a very small temperature coefficient of sensitivity ($< 0.03\%$) and has an outside diameter of 0.55 mm, which makes it ideal for such critical applications as in implanted medical devices and other invasive instruments.

A measuring system for the Fabry–Perot sensor is shown in Fig. 7.32. Light from a white-light source is coupled through a 2×2 splitter to the optical fiber that, in turn, is connected to a sensor. The sensor contains a Fabry–Perot interferometer cavity (FPI) and it reflects back light at a wavelength related to the cavity size. Now, the task is to measure the shift in a wavelength. This is accomplished by a white-light cross-correlator that contains a Fabry–Perot wedge. The wedge, in effect, is a cavity

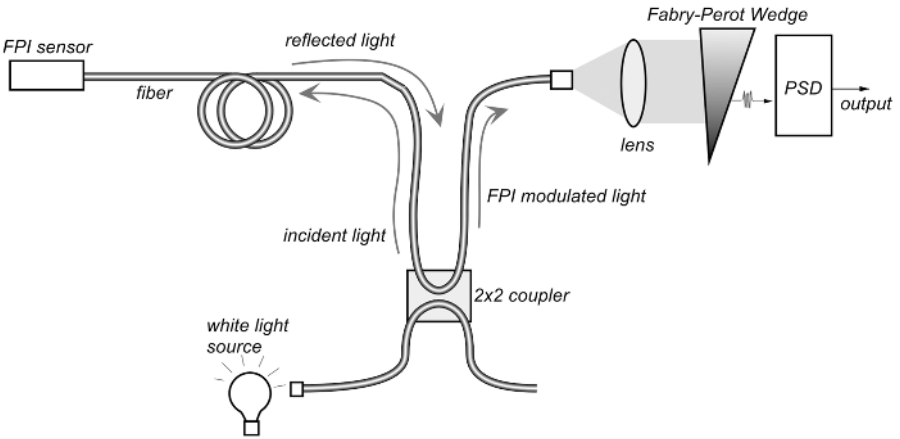


Fig. 7.32. Measuring system for the Fabry–Perot sensor. (Courtesy of Roctest. www.roctest.com.)

of a linearly variable dimension. Depending of the received wavelength, it passes light only at a specific location of the wedge. The outgoing light position at the wedge may be detected by a position-sensitive detector (PSD) that is described in detail in Section 7.5.6. The output of the detector directly relates to the input stimulus applied to the FPI sensor.

This method of sensing has the advantages of a linear response, insensitivity to the light intensity resulting from the light source or fiber transmission, versatility to measure different stimuli with the same instrument, wide dynamic range (1 : 15,000), and high resolution. In addition, the fiber-optic sensors are immune to many electromagnetic and radio-frequency interferences (EMI and RFI) and can operate reliably in harsh environment without adverse effects. For example, a FPI sensor may function inside a microwave oven.

7.5.5 Grating Sensors

An optical displacement transducer can be fabricated with two overlapping gratings which serve as a light-intensity modulator (Fig. 7.33A). The incoming pilot beam strikes the first, stationary grating which allows only about 50% of light to pass toward the second, moving grating. When the opaque sectors of the moving grating are precisely aligned with the transmitting sectors of the stationary grating, the light will be completely dimmed out. Therefore, the transmitting light beam intensity can be modulated from 0% to 50% of the pilot beam (Fig. 7.33B). The transmitted beam is focused on a sensitive surface of a photodetector, which converts light into electric current.

The full-scale displacement is equal to the size of an opaque (or clear) sector. There is a trade-off between the dynamic range of the modulator and its sensitivity; that is, for the large pitch of the grating (large sizes of the transparent and opaque

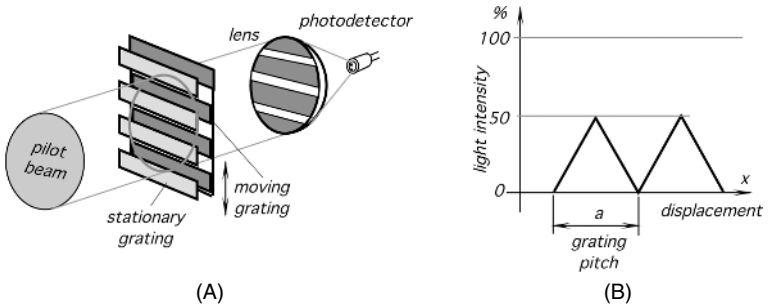


Fig. 7.33. Optical displacement sensor with grating light modulator: (A) schematic; (B) transfer function.

sectors), the sensitivity is low, but, the full-scale displacement is large. For the higher sensitivity, the grating pitch can be made very small, so that the minute movements of the grating will result in a large output signal. This type of modulator was used in a sensitive hydrophone [11] to sense displacements of a diaphragm. The grating pitch was 10 μm, which means that the full-scale displacement was 5 μm. The light source was a 2-mW He–Ne laser whose light was coupled to the grating through an optical fiber. The tests of the hydrophone have demonstrated that the device is sensitive with a dynamic range of 125 dB of pressure as referenced to 1 μPa, with a frequency response up to 1 kHz.

A grating principle of light modulation is employed in very popular rotating or linear encoders, where a moving mask (usually fabricated in the form of a disk) has transparent and opaque sections (Fig. 7.34).

The encoding disk functions as an interrupter of light beams within an optocoupler; that is, when the opaque section of the disk breaks the light beam, the detector is turned off (indicating digital ZERO), and when the light passes through a transparent section, the detector is on (indicating digital ONE). The optical encoders typically employ infrared emitters and detectors operating in the spectral range from 820 to 940 nm. The disks are made from laminated plastic and the opaque lines are produced by a photographic process. These disks are light, have low inertia and low cost and exhibit excellent resistance to shock and vibration. However, they have a limited operating temperature range. Disks for a broader temperature range are fabricated of etched metal.

There are two types of encoding disk: the incremental, which produces a transient whenever it is rotated for a pitch angle, and the absolute, whose angular position is encoded in a combination of opaque and transparent areas along the radius. The encoding can be based on any convenient digital code. The most common are the gray code, the binary, and the BCD (binary coded decimals).

The incremental encoding systems are more commonly used than the absolute systems, because of their lower cost and complexity, especially in applications where count is desirable instead of a position. When employing the incremental encoding disks, the basic sensing of movement can be made with a single optical channel (an

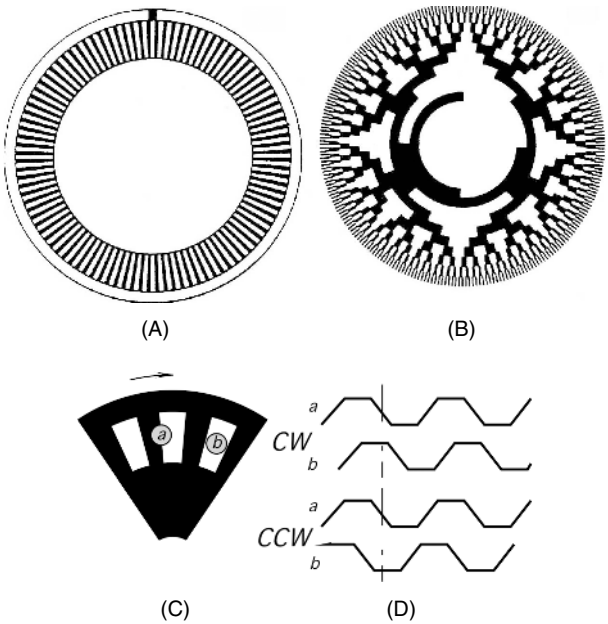


Fig. 7.34. Incremental (A) and absolute (B) optical encoding disks. When the wheel rotates clockwise (CW), the channel a signal leads b by 90° (C); when the wheel rotates counterclockwise (CCW), channel b signal leads a by 90° (D).

emitter–detector pair), whereas the speed and incremental position and the direction sensing must use two. The most commonly used approach is a quadrature sensing, where the relative position of the output signals from two optical channels are compared. The comparison provides the direction information, and either of the individual channels gives the transition signal used to derive either count or speed information (Figs. 7.31C and 7.31D).

7.5.6 Linear Optical Sensors (PSD)

For precision position measurements over short and long ranges, optical systems operating in the near infrared can be quite effective. An example is a *position-sensitive detector* (PSD) produced for precision position sensing and autofocusing in photographic and video cameras. The position measuring module is of an active type: It incorporates a light emitting diode (LED) and a photodetective PSD. The position of an object is determined by applying the principle of a triangular measurement. Figure 7.35 shows that the near-infrared LED through a collimator lens produces a narrow-angle beam ($< 2°$). The beam is a 0.7-ms-wide pulse. On striking the object, the beam is reflected back to the detector. The received low-intensity light is focused on the sensitive surface of the PSD. The PSD then generates the output signal (currents I_B and I_A), which is proportional to distance x of the light spot on its surface, from the central position. The intensity of a received beam greatly depends on the

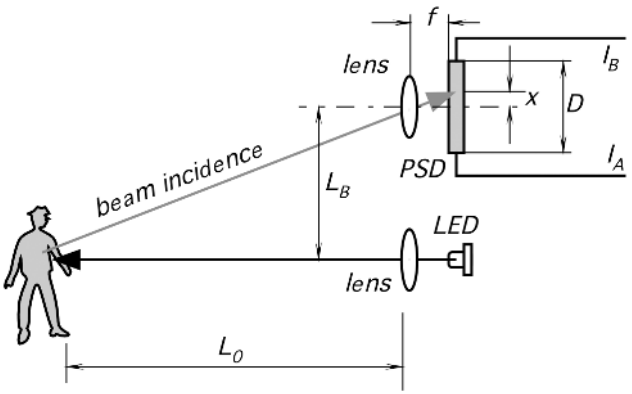


Fig. 7.35. The PSD sensor measures distance by applying a triangular principle.

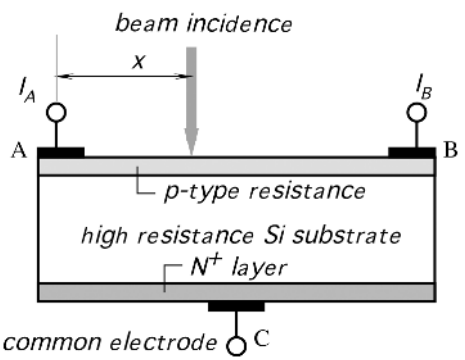


Fig. 7.36. Design of a one-dimensional PSD.

reflective properties of an object. Diffusive reflectivity in the near-infrared spectral range is close to that in the visible range; hence, the intensity of the light incident on the PSD has a great deal of variation. Nevertheless, the accuracy of measurement depends very little on the intensity of the received light.

A PSD operates on the principle of photoeffect. It makes use of a surface resistance of a silicon photodiode. Unlike MOS and CCD sensors integrating multielement photodiode arrays, the PSD has a nondiscrete sensitive area. It provides one-dimensional or two-dimensional [11] position signals on a light spot traveling over its sensitive surface. A sensor is fabricated of a piece of high-resistance silicon with two layers (p and n^+ types) built on its opposite sides (Fig. 7.36). A one-dimensional sensor has two electrodes (A and B) formed on the upper layer to provide electrical contacts to the p-type resistance. There is a common electrode (C) at the center of the bottom layer. Photoelectric effect occurs in the upper p-n junction. The distance between two upper electrodes is D, and the corresponding resistance between these two electrodes is R_D.

Let us assume that the beam incidence strikes the surface at distance x from the A electrode. Then, the corresponding resistance between that electrode and the point of incidence is respectively, R_x. The photoelectric current I_0 produced by the beam

is proportional to its intensity. That current will flow to both outputs (A and B) of the sensors in corresponding proportions to the resistances and, therefore, to the distances between the point of incidence and the electrodes:

$$I_A = I_0 \frac{R_D - R_x}{R_D} \quad \text{and} \quad I_B = I_0 \frac{R_x}{R_D}. \tag{7.9}$$

If the resistances versus distances are linear, they can be replaced with the respective distances on the surface:

$$I_A = I_0 \frac{D - x}{D} \quad \text{and} \quad I_B = I_0 \frac{c}{D}. \tag{7.10}$$

To eliminate the dependence of the photoelectric current (and of the light intensity), we can use a ratiometric technique; that is, we take the ratio of the currents,

$$P = \frac{I_A}{I_B} = \frac{D}{x} - 1, \tag{7.11}$$

which we can rewrite for a value of x:

$$x = \frac{D}{P + 1}. \tag{7.12}$$

Figure 7.35 shows geometrical relationships between various distances in the measurement system. Solving two triangles for L_0 yields

$$L_0 = f \frac{L_B}{x}, \tag{7.13}$$

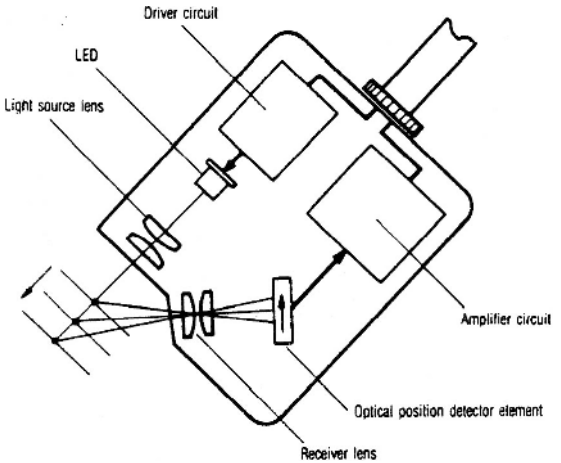
where f is the focal distance of the receiving lens. Substituting Eq. (7.12) we obtain the distance in terms of the current ratio:

$$L_0 = f \frac{L_B}{D}(P + 1) = k(P + 1), \tag{7.14}$$

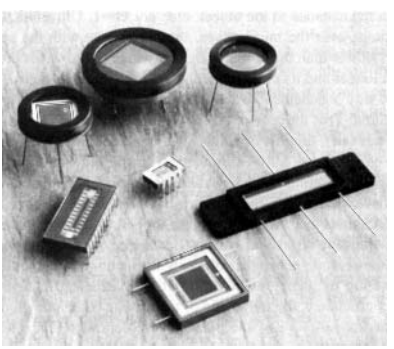
where k is called the module geometrical constant. Therefore, the distance from the module to the object linearly affects the ratio of the PSD output currents.

A similar operating principle is implemented in an industrial optical displacement sensor (Fig. 7.37) where a PSD is used for measurement of small displacements at operating distances of several centimeters. Such optical sensors are highly efficient for the on-line measurements of the height of a device (printed circuit board inspection, liquid- and solids-level control, laser torch height control, etc.), for the measurement of eccentricity of a rotating object, for thickness and precision displacement measurements, for the detection of the presence or absence of an object (medicine bottle caps), and so forth. A great advantage of an optical displacement sensor with a PSD is that its accuracy may be much greater than the accuracy of the PSD itself [12].

The PSD elements are produced of two basic types: one and two dimensional. Equivalent circuits of both are shown in Fig. 7.38. Because the equivalent circuit has a distributed capacitance and resistance, the PSD time constant varies depending



(A)



(B)

Fig. 7.37. Optical position displacement sensor (From Keyence Corp. of America, Fair Lawn, NJ.)

on the position of the light spot. In response to an input step function, a small-area PSD has rise time in the range of 1–2 μs. Its spectral response is approximately from 320 to 1100 nm; that is, the PSD covers the ultraviolet (UV), visible, and near-infrared spectral ranges. Small-area one-dimensional PSDs have sensitive surfaces ranging from 1×2 to 1×12 mm, whereas the large-area two-dimensional sensors have square areas with a side ranging from 4 to 27 mm.

7.6 Ultrasonic Sensors

For noncontact distance measurements, an active sensor which transmits some kind of a pilot signal and receives a reflected signal can be designed. The transmitted energy may be in the form of any radiation—for instance, electromagnetic in the

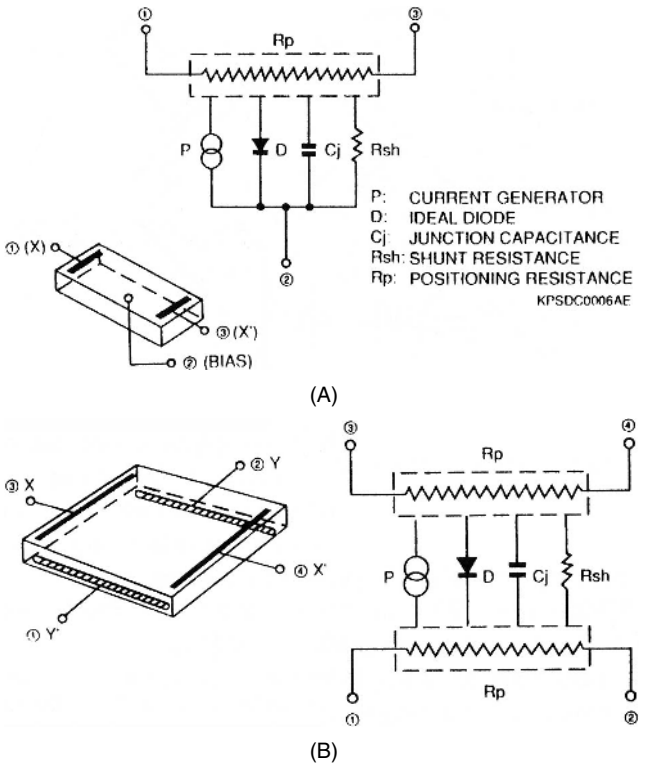


Fig. 7.38. Equivalent circuits for the (A) one- and (B) two-dimensional position-sensitive detectors. (Courtesy of Hamamatsu Photonics K.K., Japan.)

optical range (as in a PSD) electromagnetic in the microwave range, acoustic, and so forth. Transmission and reception of the ultrasonic energy is a basis for very popular ultrasonic-range meters, and velocity detectors. Ultrasonic waves are mechanical acoustic waves covering the frequency range well beyond the capabilities of human ears (i.e., over 20 kHz). However, these frequencies may be quite perceptive by smaller animals, like dogs, cats, rodents, and insects. Indeed, the ultrasonic detectors are the biological ranging devices for bats and dolphins.

When the waves are incident on an object, part of their energy is reflected. In many practical cases, the ultrasonic energy is reflected in a diffuse manner; that is, regardless of the direction from which the energy comes, it is reflected almost uniformly within a wide solid angle, which may approach 180°. If an object moves, the frequency of the reflected waves will differ from the transmitted waves. This is called the Doppler effect.[6]

[6] See Section 6.2 of Chapter 6 for the description of the Doppler effect for the microwaves. The effect is fully applicable to the propagation of any energy having a wave nature, including ultrasonic.

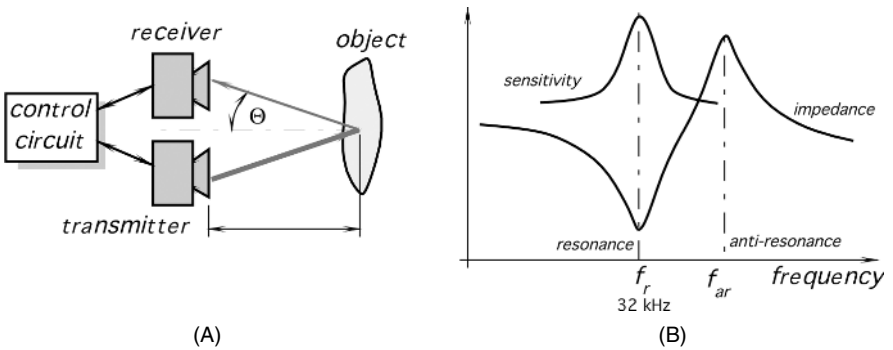


Fig. 7.39. Ultrasonic distance measurement: (A) basic arrangement; (B) impedance characteristic of a piezoelectric transducer.

The distance L_0 to the object can be calculated through the speed v of the ultrasonic waves in the media, and the angle, Θ (Fig. 7.39A):

$$L_0 = \frac{vt \cos \Theta}{2}, \tag{7.15}$$

where t is the time for the ultrasonic waves to travel to the object and back to the receiver. If a transmitter and a receiver are positioned close to each other as compared with the distance to the object, then $\cos \Theta \approx 1$. Ultrasonic waves have an obvious advantage over the microwaves: they propagate with the speed of sound, which is much slower than the speed of light at which microwaves propagate. Thus, the time t is much longer and its measurement can be accomplished easier and less expensively.

To generate any mechanical waves, including ultrasonic, the movement of a surface is required. This movement creates compression and expansion of a medium, which can be gas (air), liquids, or solids.[7] The most common type of excitation device which can generate surface movement in the ultrasonic range is a piezoelectric transducer operating in the so-called *motor* mode. The name implies that the piezoelectric device directly converts electrical energy into mechanical energy.

Figure 7.40A shows that the input voltage applied to the ceramic element causes it to flex and transmit ultrasonic waves. Because piezoelectricity is a reversible phenomenon, the ceramic generates voltage when incoming ultrasonic waves flex it. In other words, the element may work as both the transmitter and the receiver (a microphone). A typical operating frequency of the transmitting piezoelectric element is near 32 kHz. For better efficiency, the frequency of the driving oscillator should be adjusted to the resonant frequency f_r of the piezoelectric ceramic (Fig. 7.39B), where the sensitivity and efficiency of the element is best. When the measurement circuit operates in a pulsed mode, the same piezoelectric element is used for both transmission and receiving. When the system requires the continuous transmission of ultrasonic waves, separate piezoelectric elements are employed for the transmitter

[7] See Section 3.10 of Chapter 3 for description of sound waves.

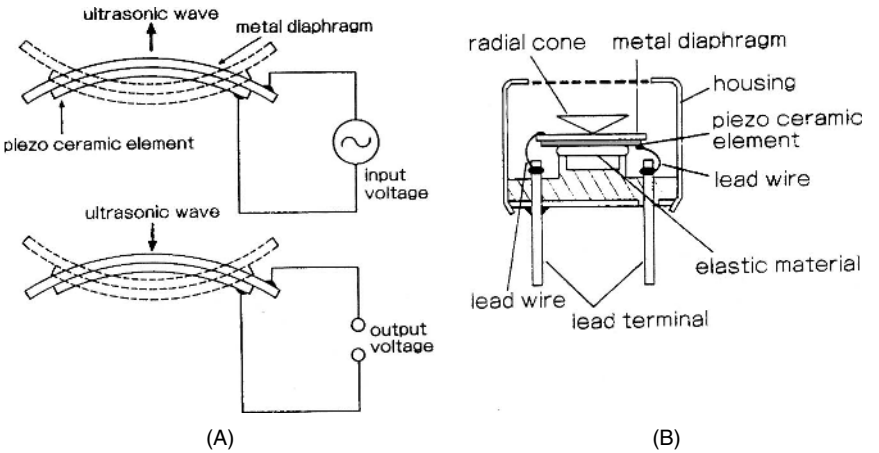


Fig. 7.40. Piezoelectric ultrasonic transducer: (A) input voltage flexes the element and transmits ultrasonic waves, whereas incoming waves produce output voltage; (B) open-aperture type of ultrasonic transducer for operation in air. (Courtesy of Nippon Ceramic, Japan.)

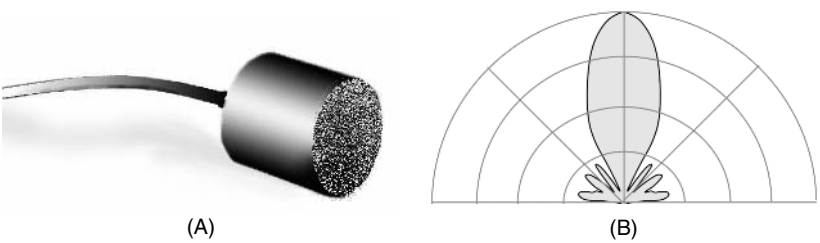


Fig. 7.41. (A) ultrasonic transducer for air. (B) directional diagram.

and receiver. A typical design of an air-operating sensor is shown in Figs. 7.40B and 7.41A. A directional sensitivity diagram (Fig. 7.41B) is important for a particular application. The narrower the diagram, the more sensitive the transducer is.

7.7 Radar Sensors

7.7.1 Micropower Impulse Radar

In 1993, Lawrence Livermore National Laboratory had developed a *micropower impulse radar* (MIR), which is a low-cost noncontact ranging sensor. The operating principle of the MIR is fundamentally the same as that of a conventional pulse radar system, but with several significant differences. The MIR (Fig. 7.42) consists of a white-noise generator whose output signal triggers a pulse generator. The pulse generator produces very short pulses with an average rate of 2 MHz $\pm$20%. Each pulse

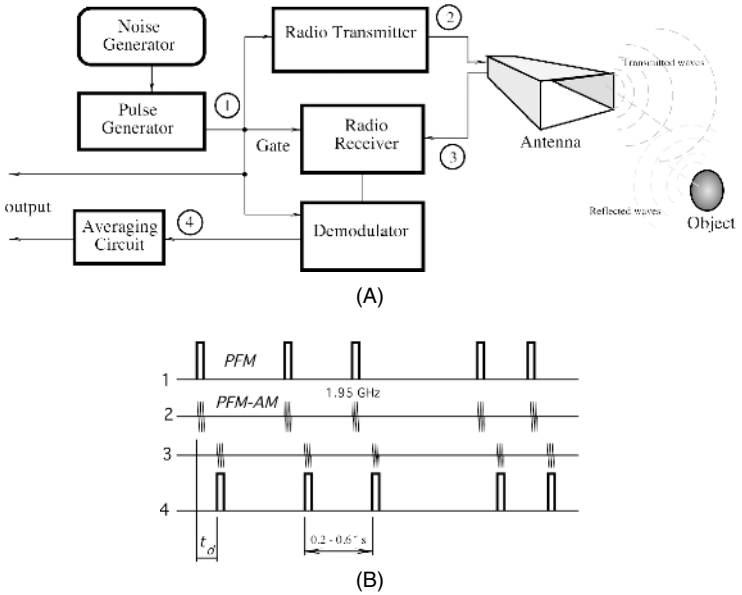
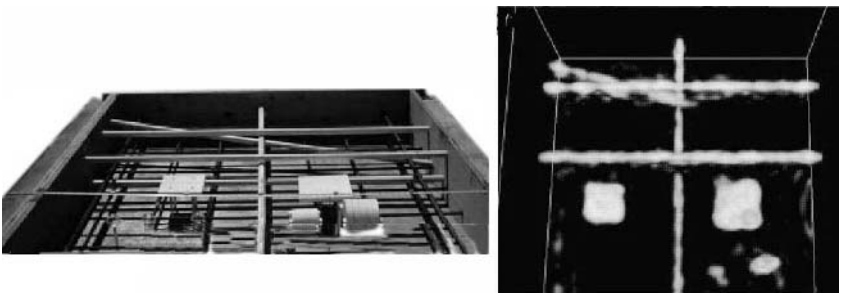


Fig. 7.42. Block diagram of micropower radar (A) and the timing diagram (B).

has a fixed short duration τ, whereas the repetition of these pulses is random, according to triggering by the noise generator. The pulses are spaced randomly with respect to one another in a Gaussian-noise like pattern. The distance between pulses range from 200 to 625 ns. It can be said that the pulses have a pulse-frequency modulation (PFM) by white noise with maximum index of 20%. In turn, the square-wave pulses cause the amplitude modulation (AM) of a radio transmitter. The modulation has a 100% depth; that is, the transmitter is turned on and off by the pulses. Such a double-step modulation is called PFM-AM.

The radio transmitter produces short bursts of high-frequency radio signal which propagate from the transmitting antenna to the surrounding space. The electromagnetic waves reflect from the objects and propagate back to the radar. The same pulse generator which modulates the transmitter gates (with a predetermined delay) the radio receiver to enable the reception of the MIR only during a specific time window. Another reason for gating the receiver is to reduce its power consumption. The reflected pulses are received and demodulated (the square-wave shape is restored from the radio signal), and the time delay with respect to the transmitted pulses is measured. The time delay is proportional to the distance D from the antenna to the object from which the radio waves are reflected: $t_d = 2Dc^{-1}$, where c is the speed of light.

The carrier frequency (center frequency) of the radio transmitter is either 1.95 or 6.5 GHz. Due to very short modulating pulses, the approximate bandwidth of the radiated signal is very wide—about 500 MHz (for a 1.95-GHz carrier). The spatial distribution of the transmitted energy is determined by the type of antenna. For a dipole antenna, it covers nearly 360°, but it may be shaped to the desired pattern



(A) (B)

Fig. 7.43. Imaging steel in concrete with MIR: (A) the internal elements of a concrete slab before pouring; (B) reconstructed 3-D MIR image of the elements embedded in the finished 30-cm-thick concrete slab.

by employing a horn, a reflector, or a lens. Because of the unpredictable modulation pattern, the wide bandwidth, and a low spectral density of the transmitted signal, the MIR system is quite immune to countermeasures and virtually is stealthy—the radiated energy is perceived by any nonsynchronous receiver as white thermal noise.

The average duty cycle of the transmitted pulses is small ($< 1\%$). Because the pulses are spaced randomly, practically any number of identical MIR systems may operate in the same space without a frequency division (i.e., they work at the same carrier frequency within the same bandwidth). There is a little chance that bursts from the interfering transmitters overlap, and if they do, the interference level is significantly reduced by the averaging circuit. Nearly 10,000 received pulses are averaged before the time delay is measured.

Other advantages of the MIR are low cost and extremely low power consumption of the radio receiver, about 12 μW. The total power consumption of the entire MIR system is near 50 μW. Two AA alkaline batteries may power it continuously for several years.

Applications for the MIR include range meters (Fig. 7.43), intrusion alarms, level detectors, vehicle ranging devices, automation systems, robotics, medical instruments, weapons, novelty products, and even toys where a relatively short range of detection is required.

7.7.2 Ground-Penetrating Radar

Civil engineering, archeology, forensic science are just a few examples of many applications of the high-frequency ground-penetrating radar (GPR). The radar operation is rather classical: It transmits radio waves and receives the reflected signal. The time delay between the transmitted and received signals is the measure of a distance to the reflecting surface. Although radars operated in air and space have ranges that reach thousands of kilometers, the GPR range at best is just several hundred meters. The radar operates at frequencies from 500 MHz to 1.5 GHz (Noggin System from Sensors & Software, Inc., Canada, www.sensoft.ca). Radio waves do not penetrate far

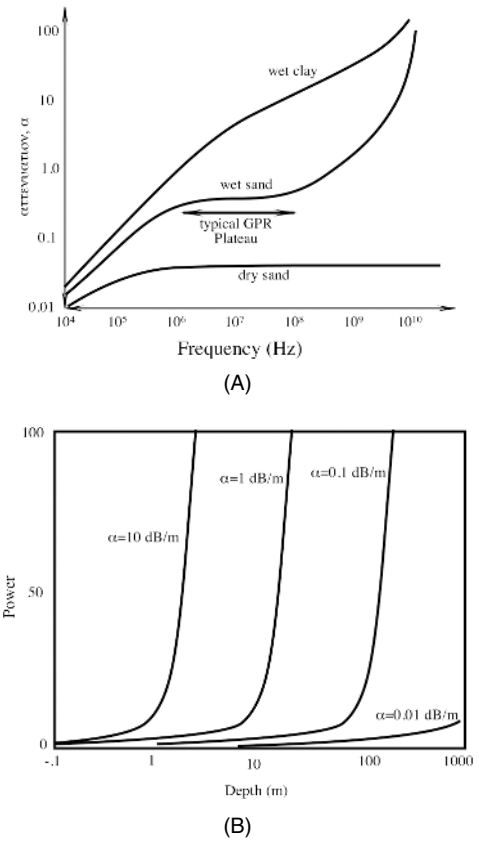
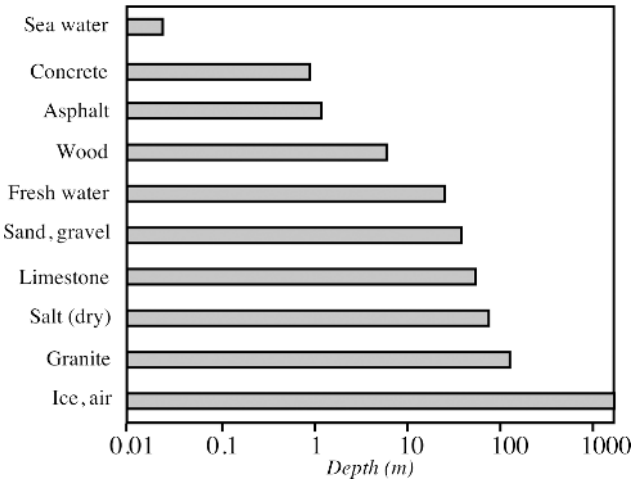


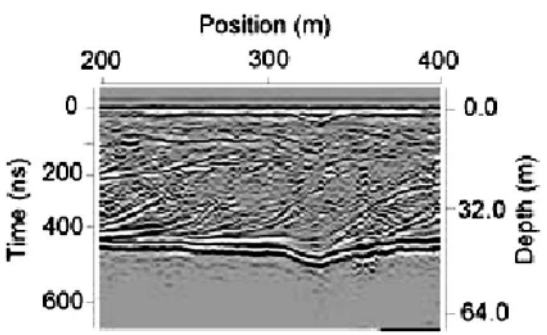
Fig. 7.44. (A) Attenuation of radio waves in different materials. Attenuation varies with excitation frequency and type of material. At low frequencies (< 1 MHz), attenuation is primarily controlled by dc conductivity. At high frequencies (> 1000 MHz), water is a strong energy absorber. (B) When attenuation limits exploration depth, power must increase exponentially with depth.

through soils, rocks, and most man-made materials such as concrete. The exponential attenuation coefficient, α, is primarily determined by the electrical conductivity of the material. In simple uniform materials, this is usually the dominant factor. In most materials, energy is also lost to scattering from material variability and to water contents. Water has two effects: First, water contains ions which contribute to bulk conductivity, and second, the water molecule absorbs electromagnetic energy at high frequencies, typically above 1000 MHz. Figure 7.44 shows that attenuation varies with excitation frequency and material. Thus, practical maximum distance increases for the dry materials (Fig. 7.45A). An example of data presented on the radar monitor is shown in Fig. 7.45B.

Lowering the frequency improves the depth of exploration because attenuation primarily increases with frequency. As the frequency decreases, however, two other fundamental aspects of the GPR measurement come into play. First, reducing the frequency results in a loss of resolution. Second, if the frequency is too low, electromagnetic fields no longer travel as waves but diffuse, which is the realm of inductive electromagnetic (EM) or eddy-current measurements.



(A)



(B)

Fig. 7.45. (A) Maximum depth for various materials; (B) graphic presentation of measured bedding in wet sand deposits. (Courtesy of Sensors & Software, Inc., Canada, www.sensoft.ca.)

7.8 Thickness and Level Sensors

In many industrial applications, the measurement of thickness of a material is essential for manufacturing, process and quality control, safety, airspace, and so forth. The methods of thickness gauging range from the optical, to ultrasonic, to x-ray. Here, we briefly review some less known methods.

7.8.1 Ablation Sensors

Ablation is dissipation of heat by melting and removal of the sacrificial protective layer during atmospheric reentry. Aerospace vehicles subjected to significant aerody-

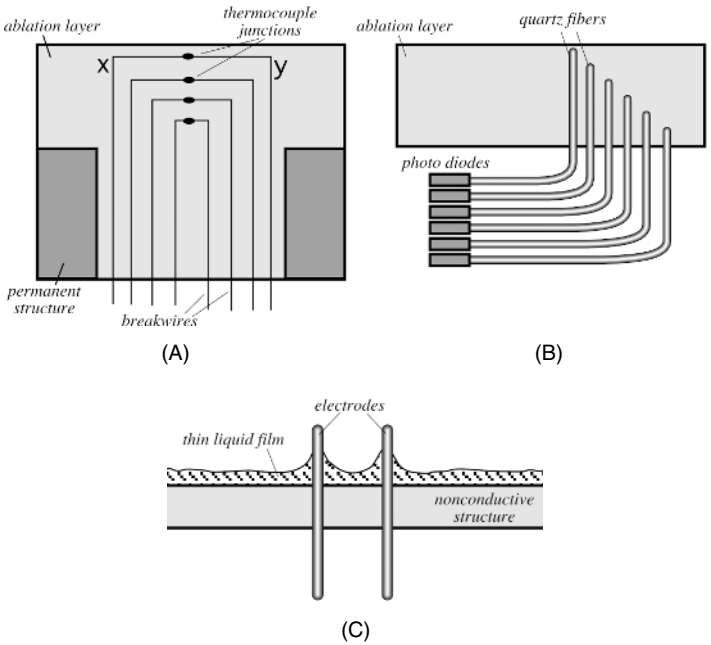
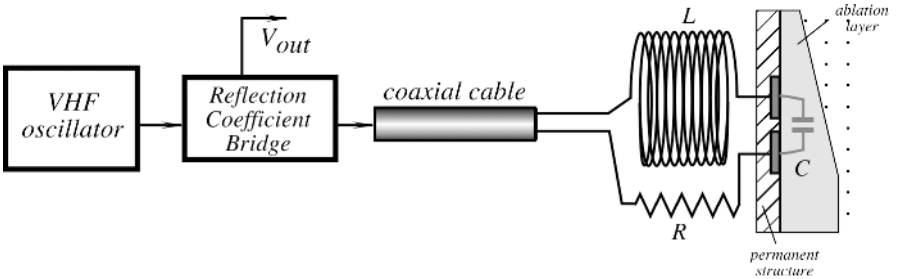


Fig. 7.46. Breakwire concept with thermocouples consisting of metals x and y (A); light pipe concept (B); and measurement of thin-film liquid by a capacitive method (C).

namic heating often rely on ablating thermal protection systems (TPSs) to keep the internal structure and equipment below critical operating temperatures. An ablating TPS undergoes chemical decomposition or phase change (or both) below the internal structure's critical temperature. Incident thermal energy is then channeled into melting, subliming, or decomposing the ablator. The ablator recession rate is directly proportional to the flux at the surface. A measure of ablator thickness is required to estimate surface heat flux. Thus, an ablation sensor is a kind of position sensor that detects position of the ablation layer's outer surface and provides a measure of the remaining thickness. The ablation sensors can be built into the ablation layer (intrusive sensors) or be noninvasive.

The intrusive sensors include the breakwire ablation gauge, radiation transducer (RAT) sensor, and light pipe [13]. The breakwire ablation gauge consists of several thin wires implanted at various known levels in an ablator. As the material progressively erodes, each successive wire is broken and results in an open circuit. Figure 7.46A illustrates this concept. In some cases [14] breakwire doubles as a thermocouple (TC) and each is situated so that no breakwire TC is directly above another. This arrangement allows an unobstructed conduction path through the ablator to each breakwire TC, including those at lower levels. Although the breakwire method provides temperature time histories until the last TC is exposed and destroyed, this method only provides recession data at a few distinct points.



(A)

(B)

Fig. 7.47. Block diagram of resonant ablation gauge (A) and prototype sensor (B).

The light pipe sensor consists of quartz fibers implanted in an ablator and terminated at known depths (Fig. 7.46B). When the TPS recedes to where a fiber terminates, light transmits down to a photodiode. This method provides recession data at distinct points only and does not provide temperature data, as the breakwire method does.

Entirely noninvasive sensor for measuring the ablation layer can be built by using the capacitive method. The sensor is made in the form of two electrodes that may have a variety of shapes [13]. The sensor is placed in series with an inductor and a resistor forming an resistive, inductive, and capacitive (RLC) termination to a waveguide (i.e., a coaxial cable). The arrangement shown in Fig. 7.47 is very similar to a transmitter–antenna configuration. The RLC termination has a resonant frequency approximated by

$$f_0 = \frac{1}{2\pi\sqrt{LC}}. \tag{7.16}$$

When electromagnetic energy at the resonant frequency is sent down the waveguide, all of the energy dissipates in the resistor. If, however, the resonant frequency of the termination changes (say, because of a change in capacitance), a fraction of the energy is reflected back toward the source. As the capacitance continues to change, the energy reflected increases. Antennas that work like this are said to be out of tune. In this situation, one could use a commercially available reflection-coefficient bridge (RCB) between the radio-frequency (RF) source and the waveguide termination. The RCB generates a dc voltage proportional to the energy reflected. Then, the antenna can be

adjusted until the bridge output voltage is a minimum and the energy transmitted is a maximum.

7.8.2 Thin-Film Sensors

Sensors for measuring the thickness of a film range from mechanical gauges, to optical, to electromagnetic and capacitive. Optical methods are limited to transparent or semitransparent films. The planar electrodes that mimic a parallel-plate capacitor produce high output; however, to be accurate, they and the sampled film must be nearly perfectly parallel, which often is not practical, especially for the surface where the film is positioned on a curvature. Thus, different types of electrode have been proposed.

An example of a simple capacitive sensor that can measure thickness of liquid film is presented in [15]. The liquid film thickness was measured via the capacitance between two small-wire probes protruding into the liquid (Fig. 7.46C). The liquid acted as a dielectric between two plates of a capacitor, with the plates being two small-wire probes. If the liquid has a different dielectric constant than air, a change in liquid level results in a change in the probe's capacitance. The capacitance changes were measured by incorporating the probe into a frequency-modulation circuit. A fixed frequency was the input to the circuit, and the output frequency depended on the probe's capacitance.

Another type of an electrode is spherical and was proposed for a dry dielectric film [16]. The capacitance is measured between the metal sphere (a stainless-steel ball having a diameter between 3 and 4 mm) and a conductive base (Fig. 7.48A). To minimize a fringing effect, the ball is surrounded by a driven shield that helps in directing the electric field only toward the base electrode through the film.

7.8.3 Liquid-Level Sensors

There are many ways to detect levels of liquids. They include the use of the resistive (see Fig. 7.1B), optical (see Fig. 7.28), magnetic (see Fig. 7.24), and capacitive (see Fig. 3.8 of Chapter 3) sensors. The choice of a particular sensor depends on many factors, but probably the defining factor is the type of a liquid. One of the most challenging is liquid gases, especially liquid helium, which has a low density and low dielectric constant, not mentioning its storage in the enclosed Dewar bottles at a cryogenic temperature. Is such difficult cases, a transmission-line sensor may be quite efficient. The sensor operates on a principle that is similar to the one that was described for ablation sensing (Fig. 7.47). For detecting the liquid levels, the transmission-line sensor may be constructed as shown in Fig. 7.49.

The probe resembles a capacitive-level sensor shown in Fig. 3.8 of Chapter 3; however, its operation does not rely on the liquid dielectric constant, as is the case in Fig. 3.8. The probe looks like a long tube with an inner electrode surrounded by the outer cylindrical electrode. The probe is immersed into liquid, which may freely fill the space between the electrodes. The electrodes are fed with a high-frequency

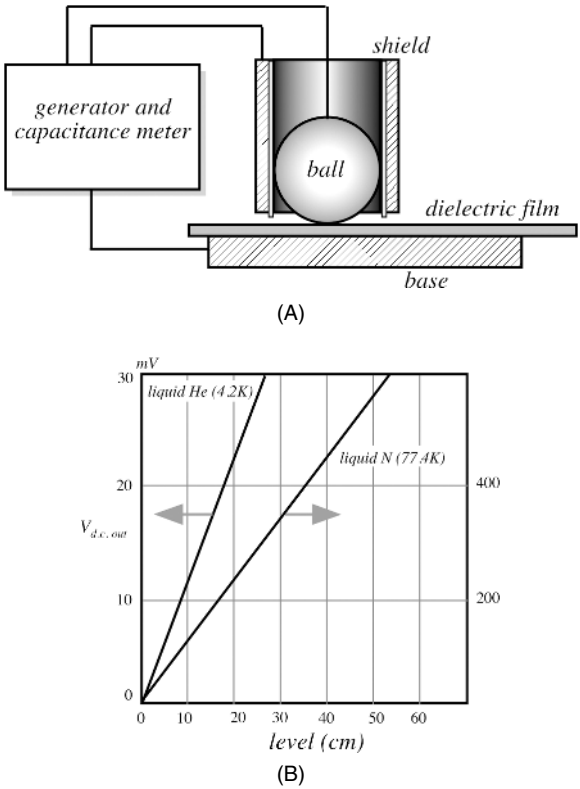
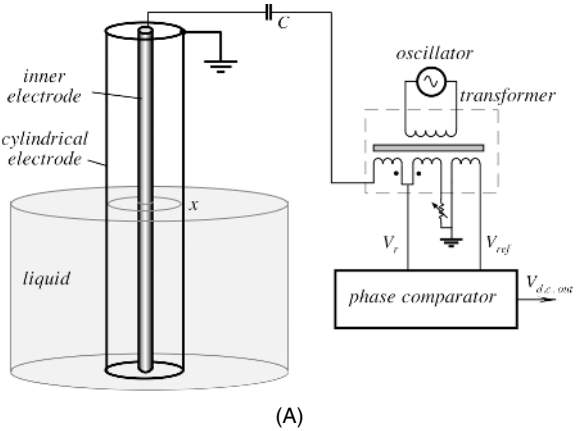
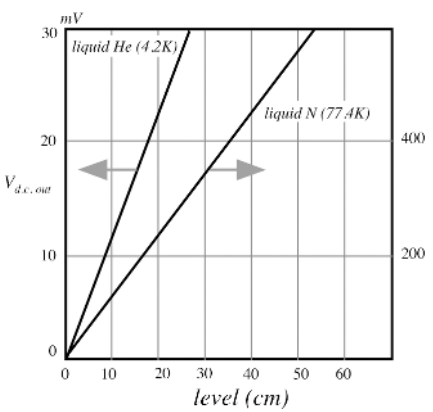


Fig. 7.48. Dry dielectric film capacitive sensor (A) and shape of transfer function (B). (Adapted from Ref. [16].)

signal (about 10 MHz). A length of the probe can be any practical wavelength but for a linear response, it is advisable to keep it less than $(1/4)\lambda$ [17]. The high-frequency signal propagates along the transmission line that is formed by the two electrodes. The liquid fills the space between the electrodes up to a particular level x. Because the dielectric constant of a liquid is different from its vapor, the properties of the transmission line depend on the position of the borderline between liquid and vapor (in other words, on the liquid level). The high-frequency signal is partially reflected from the liquid–vapor borderline and propagates back toward the upper portion of the sensor. To some degree, it resembles radar that sends a pilot signal and receives the reflection. By measuring a phase shift between the transmitted and reflected signals, the position of the borderline can be computed. The phase-shift measurement is resolved by a phase comparator that produces a dc voltage at its output. A higher dielectric constant produces a better reflection and, thus, the sensitivity of the sensor improves accordingly (Fig. 7.49B).



(A)



(B)

Fig. 7.49. Transmission-line probe (A) and transfer functions (B). (Adapted from Ref. [17].)

References

1. Kato, H., Kojima, M., Gattoh, M., Okumura, Y., and Morinaga, S. Photoelectric inclination sensor and its application to the measurement of the shapes of 3-D objects. *IEEE Trans. Instrum. Meas.* 40(6), 1021–1026, 1991.
2. Barker, M. J. and Colclough, M. S. A two-dimensional capacitive position transducer with rotation output. *Rev. Sci. Instrum.*, 68(8), 3238–3240, 1997.
3. Peters, R.D. U.S. Patent 5,461,319, 1995.
4. De Silva, C. W. *Control Sensors and Actuators*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
5. *Linear Application Handbook*, Linear Technology, 1990.
6. *Hall Effect IC Applications*, Sprague, 1986.
7. B. Halg, A silicon pressure sensor with a low-cost contactless interferometric optical readout. *Sensors Actuators A* 30, 225–229, 1992.

8. Dakin, J. P., Wade, C. A. and Withers, P. B. An optical fiber pressure sensor, *SPIE Fiber Optics '87: Fifth International Conference on Fiber Optics and Optoelectronics*, Bellingham, 1987, pp. 194–201.
9. Lee, C. E. and Taylor, H. F. Fiber-optic Fabry–Perot temperature sensor using a low-coherence light source. *J. Lightwave Technol.* 129–134, 1991.
10. Wolthuis, R. A., Mitchell, G. L., Saaski, E., Hartl, J. C. and Afromowitz, M. A. Development of medical pressure and temperature sensors employing optical spectrum modulation. *IEEE Trans. Biomed. Eng.* 38, 974–980, 1991.
11. Spillman, W.B., Jr. Multimode fiber-optic hydrophone based on a Schlieren technique. *Appl. Opt.* 20, 465, 1981.
12. van Drecht, J. and Meijer, G.C.M. Concepts for the design of smart sensors and smart signal processors and their applications to PSD displacement transducers. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*. IEEE, New York, 1991, pp. 475–478.
13. Noffz, G. K. and Bowman, M. P. Design and laboratory validation of a capacitive sensor for measuring the recession of a thin-layered ablator. NASA Technical Memorandum 4777, 1996.
14. *In-Depth Ablative Plug Transducers*, Hycal Engineering, El Monte, CA, 1992.
15. Brown, R. C., Andreussi, P., and Zanelli, S. The use of wire probes for the measurement of liquid film thickness in annular gas-liquid flows. *Can. J. Chem. Eng.*, 56, 754–757, 1978.
16. Graham, J., Kryzeminski, M., and Popovic, Z. Capacitance based scanner for thickness mapping of thin dielectric films. *Rev. Sci. Intrum.*, 71(5), 2219–2223, 2000.
17. Brusch, L., Delfitto, G., and Mistura, G. Level meter for dielectric liquids. *Rev. Sci. Instrum.* 70(2), 1999.

This page intentionally left blank

8

Velocity and Acceleration

Acceleration is a dynamic characteristic of an object, because, according to Newton's second law, it essentially requires application of a force. In effect, the position, velocity, and acceleration are all related: Velocity is a first derivative of position and acceleration is the second derivative. However, in a noisy environment, taking derivatives may result in extremely high errors, even if complex and sophisticated signal conditioning circuits are employed. Therefore, velocity and acceleration are not derived from the position detectors, but rather measured by special sensors. As a rule of thumb, in low-frequency applications (having a bandwidth on the order of 1 Hz), position and displacement measurements generally provide good accuracy. In the intermediate-frequency applications (less than 1 kHz), velocity measurement is usually favored. In measuring high-frequency motions with appreciable noise levels, acceleration measurement is preferred.

Velocity (speed or rate of motion) may be linear or angular; that is, it shows how fast an object moves along a straight line or how fast it rotates. The measure of velocity depends on the scale of an object and may be expressed, say, in millimeters per second or miles per hour. Currently, the speed of a large object, especially of a land or water vehicle, may be very efficiently determined by a GPS (Geo Positioning System) that operates by receiving radio signals from a number of the Earth's satellites and by computing the time delay of signals received from one satellite as compared with the other. When the position of a vehicle is determined with a periodic rate, computation of its velocity is no problem. For smaller objects and shorter distances, GPS is not a solution. Detecting the velocity for such objects requires different references. A basic idea behind many sensors for the transduction of velocity or acceleration is a measurement of the displacement of an object with respect to some reference object which, in many cases, is an integral part of the sensor. *Displacement* here is a keyword. Many velocity or acceleration sensors contain components which are sensitive to a displacement. Thus, the position and displacement sensors described are the integral parts of the velocity sensors and accelerometers. In some instances, however, velocity sensors and accelerometers do not use an intermediate displacement transducer because their motions may be directly converted into electrical signals. For example, moving a magnet though a coil of wire will induce a voltage in the coil according

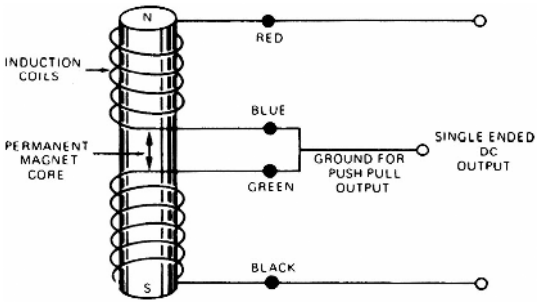


Fig. 8.1. Operating principle of an electromagnetic velocity sensor. (Courtesy of Trans-Tek, Inc., Ellington, CT.)

to Faraday's law. This voltage is proportional to the magnet's *velocity* and the field strength [Eq. (3.39) of Chapter 3]. Linear velocity transducers use this principle of magnetic induction, with a permanent magnet and a fixed geometry coil, so the output voltage of the coil is directly proportional to the magnet's relative velocity over its working range.

In the velocity sensor, both ends of the magnet are inside the coil. With a single coil, this would give a zero output because the voltage generated by one end of the magnet would cancel the voltage generated by the other end. To overcome this limitation, the coil is divided into two sections. The north pole of the magnet induces a current in one coil, and the south pole induces a current in the other coil (Fig. 8.1). The two coils are connected in a series-opposite direction to obtain an output proportional to the magnet's velocity. The maximum detectable velocity depends primarily on the input stages of the interface electronic circuit. The minimum detectable velocity depends on the noise floor and especially of transmitted noise from nearby high-ac-current equipment. Typical specifications of an electromagnetic sensor are given in Table 8.1. This design is very similar to a linear variable differential transformer (LVDT) position sensor (Section 7.4 of Chapter 7), except that the LVDT is an active sensor with a moving ferromagnetic core, whereas the velocity sensor is a passive device with a moving permanent magnet; that is, this sensor is a current-generating device

Table 8.1. Specification Ranges of Electromagnetic Velocity Sensors

Characteristic	Value
Magnet core displacement (in.)	0.5–24
Sensitivity (mV/in./s)	35–500
Coil resistance (kΩ)	2–45
Coil inductance (H)	0.06–7.5
Frequency response (Hz) (at load >100 times the coil resistance)	500–1500
Weight (g)	20–1500

Source: Courtesy of Trans-Tek, Inc., Ellington, CT.

which does not need an excitation signal. Naturally, linear velocity sensors detect velocity along a distance that is limited by the size of the sensor; therefore, in most cases, these sensors measure vibration velocity. An angular version of the same sensor can measure rotation rate continuously for any number of turns.

8.1 Accelerometer Characteristics

Vibration is a dynamic mechanical phenomenon which involves periodic oscillatory motion around a reference position. In some cases (shock analysis, linear acceleration, etc.), the oscillating aspect may be missing, but the measurement and design of the sensor remains the same. An accelerometer can be specified as a single-degree-of-freedom device which has some type of seismic mass (sometimes called *proof mass*), a springlike supporting system, and a frame structure with damping properties (Fig. 3.48A of Chapter 3).

A mathematical model of an accelerometer is represented by Eq. (3.156) of Chapter 3. To solve the equation, it is convenient to use the Laplace transformation, which yields

$$Ms^2X(s) + bsX(s) + kX(s) = -MA(s), \tag{8.1}$$

where $X(s)$ and $A(s)$ are the Laplace transforms of $x(t)$ and d^2y/dt^2, respectively.[1] Solving for $X(s)$, we obtain

$$X(s) = -\frac{MA(s)}{Ms^2 + bs + k}. \tag{8.2}$$

We introduce a conventional variable $\omega_0 = \sqrt{k/M}$, and $2\zeta\omega_0 = b/M$, then Eq. (8.2) can be expressed as

$$X(s) = -\frac{A(s)}{s^2 + 2\xi\omega_0 s + \omega_0^2}. \tag{8.3}$$

The value of ω_0 represents the accelerometer's angular natural frequency and ζ is the normalized damping coefficient. Let us set

$$G(s) = -\frac{1}{s^2 + 2\xi\omega_0 s + \omega_0^2}; \tag{8.4}$$

then, Eq. (8.3) becomes $X(s) = G(s)A(s)$ and the solution can be expressed in terms of the inverse Laplace transform operator as

$$x(t) = L^{-1}\{G(s)A(s)\}, \tag{8.5}$$

which, from the convolution theorem for the Laplace transform, can be expressed as

$$x(t) = \int_0^t g(t-\tau)a(\tau)\,d\tau, \tag{8.6}$$

[1] d^2y/dt^2 is the input acceleration of the accelerometer body.

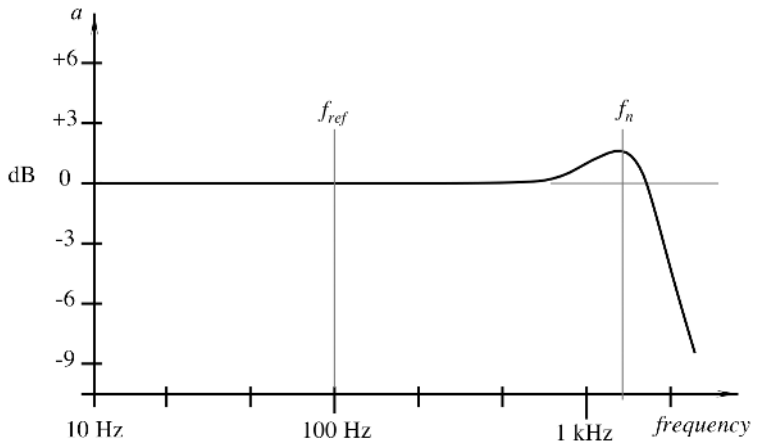


Fig. 8.2. A frequency response of an accelerometer. f_n is a natural frequency and f_{ref} is the reference frequency.

where a is the time-dependent impulse of the accelerometer body and $g(t)$ is the inverse transform $L^{-1}\{G(s)\}$. If we set $\omega = \omega_0\sqrt{1-\zeta^2}$, then Eq. (8.6) has two solutions. One is for the underdamped mode ($\zeta < 1$),

$$x(t) = \int_0^t -\frac{1}{\omega} e^{-\zeta\omega_0(t-\tau)} \sin\omega(t-\tau)a(t)\,d\tau, \tag{8.7}$$

whereas for the overdamped mode ($\zeta > 1$),

$$x(t) = \int_0^t -\frac{1}{\omega} e^{-\zeta\omega_0(t-\tau)} \sinh\omega(t-\tau)a(t)\,d\tau, \tag{8.8}$$

where $\omega = \omega_0\sqrt{\zeta^2-1}$. The above solutions can be evaluated for different acceleration inputs applied to the accelerometer base [1].

A correctly designed, installed, and calibrated accelerometer should have one clearly identifiable resonant (natural) frequency and a flat frequency response at which the most accurate measurement can be made (Fig. 8.2). Within this flat region, as the vibrating frequency changes, the output of the sensor will correctly reflect the change without multiplying the signal by any variations in the frequency characteristic of the accelerometer. Viscous damping is used in many accelerometers to improve the useful frequency range by limiting the effects of the resonant. As a damping medium, silicone oil is used quite often.

When calibrated, several characteristics of an accelerometer should be determined:

1. Sensitivity is the ratio of an electrical output to the mechanical input. It is usually expressed in terms of volts per unit of acceleration under the specified conditions. For instance, the sensitivity may be specified as 1 V/g (unit of acceleration: $g = 9.80665$ m/s^2 at sea level, 45° latitude). The sensitivity is typically measured

at a single reference frequency of a sine-wave shape. In the United States, it is 100 Hz, and in most European countries, it is 160 Hz.[2]
2. Frequency response is the output's signal over a range of frequencies where the sensor should be operating. It is specified with respect to a reference frequency, which is where the sensitivity is specified.
3. Resonant frequency in an undamped sensor shows as a clearly defined peak that can be 3–4 dB higher than the response at the reference frequency. In a near-critically damped device, the resonant may not be clearly visible; therefore, the phase shift is measured. At the resonant frequency, it is 180° of that at the reference frequency.
4. Zero stimulus output (for capacitive and piezoresistive sensors) is specified for the position of the sensor where its sensitive (active) axis is perpendicular to Earth's gravity; that is, in the sensors which have a dc component in the output signal, the gravitational effect should be eliminated before the output, as no mechanical input is determined.
5. Linearity of the accelerometer is specified over the dynamic range of the input signals.

When specifying an accelerometer for a particular application, one should answer a number questions:

1. What is the anticipated magnitude of vibration or linear acceleration?
2. What is the operating temperature and how fast can the ambient temperature change?
3. What is the anticipated frequency range?
4. What linearity and accuracy are required?
5. What is the maximum tolerable size?
6. What kind of power supply is available?
7. Are any corrosive chemicals or high moisture present?
8. What is an anticipated overshock?
9. Are intense acoustic, electromagnetic, or electrostatic fields present?
10. Is the machinery grounded?

8.2 Capacitive Accelerometers

An accelerometer requires a special component whose movement lags behind that of the accelerometer's housing, which is coupled to the object under study. Then, a displacement transducer can be employed to generate an electrical signal as a function, or proof of the acceleration. This component is usually called either a *seismic* or an *inertial* mass. Regardless of the sensors' design or the conversion technique, an ultimate goal of the measurement is the *detection of the mass displacement* with respect to the accelerometer housing. Hence, any suitable displacement transducer capable of measuring microscopic movements under strong vibrations or linear acceleration

[2] These frequencies are chosen because they are removed from the power-line frequencies and their harmonics.

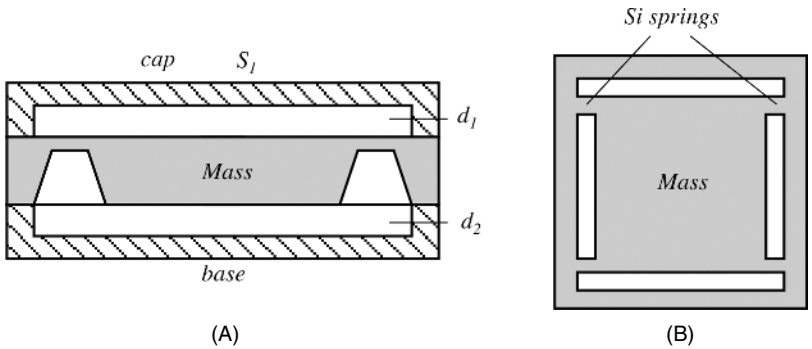


Fig. 8.3. Capacitive accelerometer with a differential capacitor: (A) side cross-sectional view; (B) top view of a seismic mass supported by four silicon springs.

can be used in an accelerometer. A capacitive displacement conversion is one of the proven and reliable methods. A capacitive-acceleration sensor essentially contains at least two components; the first is a "stationary" plate (i.e., connected to the housing) and the other is a plate attached to the inertial mass which is free to move inside the housing. These plates form a capacitor whose value is a function of a distance d between the plates [Eq. (3.23) of Chapter 3]. It is said that the capacitor value is modulated by the acceleration. A maximum displacement which is measured by the capacitive accelerometer rarely exceeds 20 μm. Hence, such a small displacement requires a reliable compensation of drifts and various interferences. This is usually accomplished by the use of a differential technique, where an additional capacitor is formed in the same structure. The value of the second capacitor must be close to that of the first, and it should be subjected to changes with a 180° phase shift. Then, an acceleration can be represented by a difference in values between the two capacitors.

Figure 8.3A shows a cross-sectional diagram of a capacitive accelerometer where an internal mass is sandwiched between the upper cap and the base [2]. The mass is supported by four silicon springs (Fig. 8.3B). The upper plate and the base are separated from it by respective distances d_1 and d_2. All three parts are micromachined from a silicon wafer. Figure 8.4 is a simplified circuit diagram for a capacitance-to-voltage converter, which in many respects is similar to the circuit of Fig. 5.52 of Chapter 5.

A parallel-plate capacitor C_{mc} between the mass and the cap electrodes has a plate area S_1. The plate spacing d_1 can be reduced by an amount Δ when the mass moves toward the upper plate. A second capacitor C_{mb} having a different plate area S_2 appears between the mass and the base. When mass moves toward the upper plate and away from the base, the spacing d_2 increases by Δ. The value of Δ is equal to the mechanical force F_m acting on the mass divided by the spring constant k of the silicon springs:

$$\Delta = \frac{F_m}{k}. \tag{8.9}$$

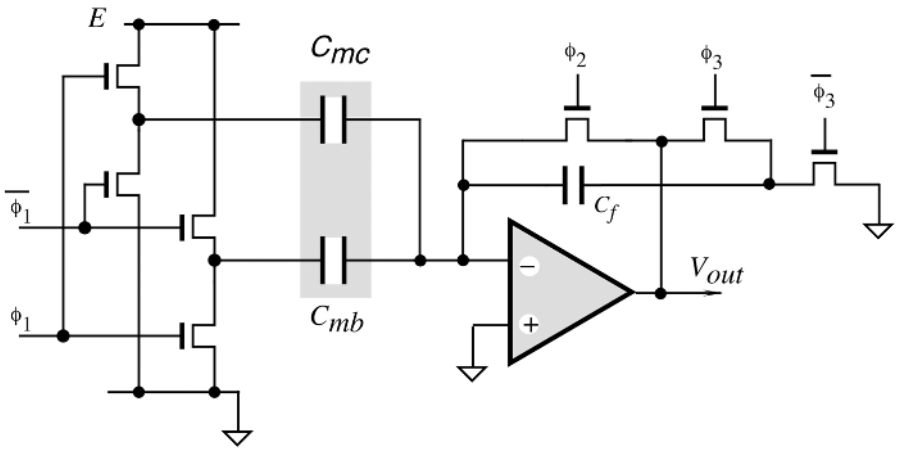


Fig. 8.4. Circuit diagram of a capacitance-to-voltage conversion suitable for an integration on silicon.

Strictly speaking, the accelerometer equivalent circuit is valid only when electrostatic forces do not affect the mass position (i.e., when the capacitors depend linearly on F_m) [3]. When an accelerometer serves as the input capacitor to a switched-capacitor summing amplifier, the output voltage depends on the value of the capacitors and, subsequently, on force:

$$V_{\text{out}} = 2E\frac{C_{\text{mc}} - C_{\text{mb}}}{C_f}. \tag{8.10}$$

Equation (8.10) is true for small changes in the sensor's capacitances. The accelerometer output is also a function of temperature and a capacitive mismatch. It is advisable that it be calibrated over an entire temperature range and an appropriate correction is made during the signal processing. Another effective method of assuring high stability is to design self-calibrating systems which make use of electrostatic forces appearing in the accelerometer assembly when a high voltage is applied to either a cap or base electrode.

8.3 Piezoresistive Accelerometers

As a sensing element, a piezoresistive accelerometer incorporates strain gauges, which measure strain in mass-supporting springs. The strain can be directly correlated with the magnitude and rate of mass displacement and, subsequently, with an acceleration. These devices can sense accelerations within a broad frequency range: from dc up to 13 kHz. With a proper design, they can withstand overshock up to 10,000g. Naturally, a dynamic range (span) is somewhat narrower ($\pm 1000g$ with error less than 1%). The overshock is a critical specification for many applications. However, piezoresistive

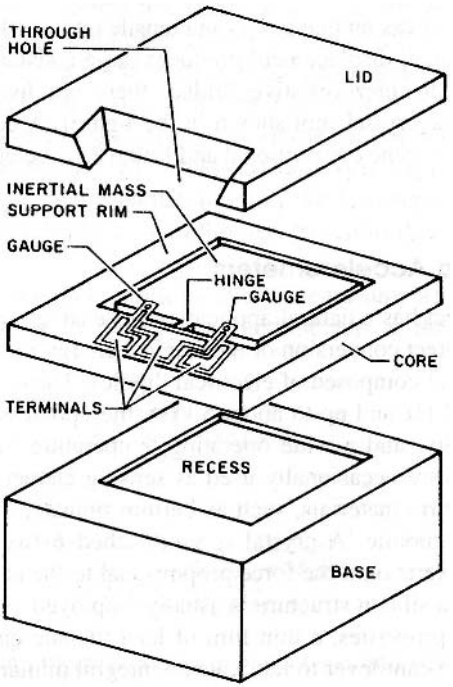


Fig. 8.5. Exposed view of a piezoresistive accelerometer.

accelerometers with discrete, epoxy-bonded strain gauges tend to have undesirable output temperature coefficients. Because they are manufactured separately, the gauges require individual thermal testing and parameter matching. This difficulty is virtually eliminated in modern sensors, which use the micromachining technology of silicon wafers.

An example of a wide-dynamic-range solid-state accelerometer is shown in Fig. 8.5. It was developed by Endevco/Allied Signal Aerospace Co. (Sunnyvale, CA). The microsensor is fabricated from three layers of silicon. The inner layer, or the core, consists of an inertial mass and the elastic hinge. The mass is suspended inside an etched rim on the hinge, which has piezoresistive gauges on either side. The gauges detect motion about the hinge. The outer two layers, the base and the lid, protect the moving parts from the external contamination. Both parts have recesses to allow the inertial mass to move freely [4]. Several important features are incorporated into the sensor. One is that the sensitive axis lies in the plane of the silicon wafer, as opposed to many other designs where the axis is perpendicular to the wafer. Mechanical integrity and reliably are assured by the fabrication of all of the components of the sensor from a single silicon crystal.

When acceleration is applied along the sensitive axis, the inertial mass rotates around the hinge. The gauges on both sides of the hinge allow rotation of the mass to create compressive stress on one gauge and tensile on the other. Because gauges are very short, even the small displacement produces large resistance changes. To trim the zero balance of the piezoresistive bridge, there are five trimming resistors positioned on the same crystal (not shown in Fig. 8.5).

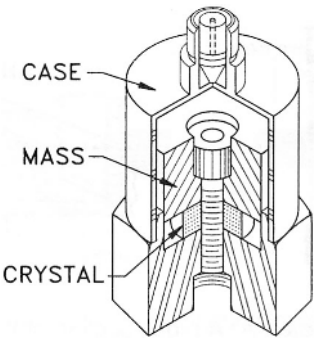


Fig. 8.6. Basic schematic of a piezoelectric accelerometer. Acceleration of the case moves it relative to the mass, which exerts a force on the crystal. The output is directly proportional to the acceleration or vibration level.

8.4 Piezoelectric Accelerometers

The piezoelectric effect (do not confuse it with a piezoresistive effect) has a natural application in sensing vibration and acceleration. The effect is a direct conversion of mechanical energy into electrical energy (Section 3.6 of Chapter 3) in a crystalline material composed of electrical dipoles. These sensors operate from frequencies as low as 2 Hz and up to about 5 kHz; they posses good off-axis noise rejection, high linearity, and a wide operating temperature range (up to 120°C). Although quartz crystals are occasionally used as sensing elements, the most popular are ceramic piezoelectric materials, such as barium titanate, lead zirconite titanate (PZT), and lead metaniobite. A crystal is sandwiched between the case and the seismic mass which exerts a force proportional to the acceleration on it (Fig. 8.6). In miniature sensors, a silicon structure is usually employed. Because silicon does not possess piezoelectric properties, a thin film of lead titanate can be deposited on a micromachined silicon cantilever to fabricate an integral miniature sensor. For good frequency characteristics, a piezoelectric signal is amplified by a charge-to-voltage or current-to-voltage converter which usually is built into the same housing as the piezoelectric crystal.

8.5 Thermal Accelerometers

8.5.1 Heated-Plate Accelerometer

Because the basic idea behind an accelerometer is a measurement of the movement of seismic mass, a fundamental formula of heat transfer can be used for that measurement [see Eq. (3.125) of Chapter 3]. A thermal accelerometer, as any other accelerometer, contains a seismic mass suspended by a thin cantilever and positioned in close proximity to a heat sink or between two heat sinks (Fig. 8.7) [5]. The mass and the cantilever structure are fabricated using micromachine technology. The space between these components is filled with a thermally conductive gas. The mass is heated by a surface or imbedded heater to a defined temperature T_1. Under the no-acceleration conditions a thermal equilibrium is established between the mass and the heat sinks: the amounts of heat q_1 and q_2 conducted to the heat sinks through gas from the mass is a function of distances M_1 and M_2.

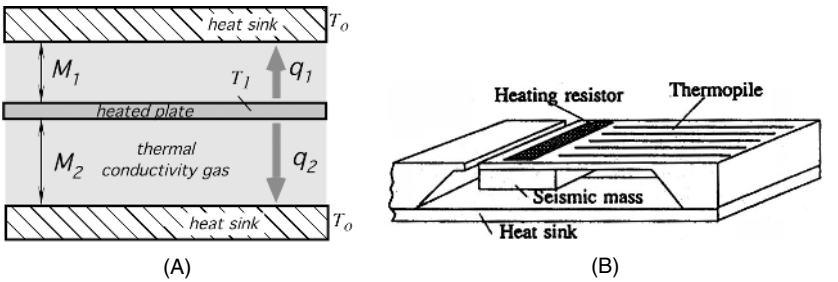


Fig. 8.7. Thermal accelerometer: (A) cross section of the heated part; (B) an accelerometer design (shown without the roof) (adapted from [5])

The temperature at any point in the cantilever beam supporting the seismic mass[3] depends on its distance from the support x and the gaps at the heat sinks. It can be found from

$$\frac{d^2T}{dx^2} - \lambda^2 T = 0, \tag{8.11}$$

where

$$\lambda = \sqrt{\frac{K_g(M_1 + M_2)}{L_{\text{si}} D M_1 M_2}}, \tag{8.12}$$

where K_g and K_{si} are thermal conductivities of gas and silicon, respectively, and D is the thickness of a cantilever beam. For a boundary conditions, where the heat sink temperature is zero, a solution of Eq. (8.11) is

$$T(x) = \frac{P \sinh(\lambda x)}{W D K_{\text{si}} \lambda \cosh(\lambda L)}, \tag{8.13}$$

where W and L is the width and length of the beam, respectively, and P is the thermal power. To measure that temperature, a temperature sensor can be deposited on the beam. It can be done by integrating silicone diodes into the beam,[4] or by forming serially connected thermocouples (a thermopile) on the beam surface. Eventually, the measured beam temperature in the form of an electrical signal is a measure of acceleration. The sensitivity of a thermal accelerometer (about 1% of change in the output signal per g) is somewhat smaller than that of the capacitive or piezoelectric types; however, it is much less susceptible to such interferences as ambient temperature or electromagnetic and electrostatic noise.

8.5.2 Heated-Gas Accelerometer

Another interesting accelerometer uses gas as a seismic mass. The heated-gas accelerometer (HGA) was developed by MEMSIC Corporation (www.memsic.com). It is fabricated on a micromachined CMOS chip and is a complete biaxial motion measurement system. The principle of operation of the device is based on heat transfer by

[3] Here, we assume steady-state conditions and neglect radiative and convective heat transfers.

[4] See Chapter 16 for a description of a Si diode as a temperature sensor.

forced convection. As described in Chapter 3, heat can be transferred by conduction, convection, and radiation. Convection can be natural (caused by gravity) or forced (by applying an artificial external force, like that produced by a blower). In a HGA, such force is produced by acceleration. The sensor measures the internal changes in heat transfer of the trapped gas. The sensor is functionally equivalent to traditional inertial mass accelerometers. The inertial mass in the sensor is gas that is thermally nonhomogeneous. The gaseous inertial mass provides some advantages over the use of the traditional solid inertial mass. The most important advantage is a shock survival up to 50,000g, leading to significantly lower failure rates.

The sensor contains a micromachined plate adjacent to a sealed cavity filled with gas. The plate has an etched cavity (trench). A single heat source, centered in the silicon chip, is suspended across the trench (Fig. 8.8). Equally spaced are four temperature sensors that are aluminum/polysilicon thermopiles (TP) (i.e., serially connected thermocouples). The TPs are located equidistant on all four sides of the heat source (dual axis). Note that a TP measures only a temperature gradient, so that the left and right thermopiles in fact is a single TP, where the left portion is the location of "cold" junctions and the right portion is that of "hot" junctions (see Section 16.2 of Chapter 16 for the operating principle of a thermocouple). A thermopile instead of a thermocouple is used for a sole purpose: to increase the electrical output signal. Another pair of junctions is used for measuring a thermal gradient along the y axis.

Under zero acceleration, a temperature distribution across the gas cavity is symmetrical about the heat source, so that the temperature is the same at all four TP junctions, causing each pair to output zero voltage. The heater is warmed to a temperature that is well above ambient and typically is near 200°C. Figure 8.8A shows two

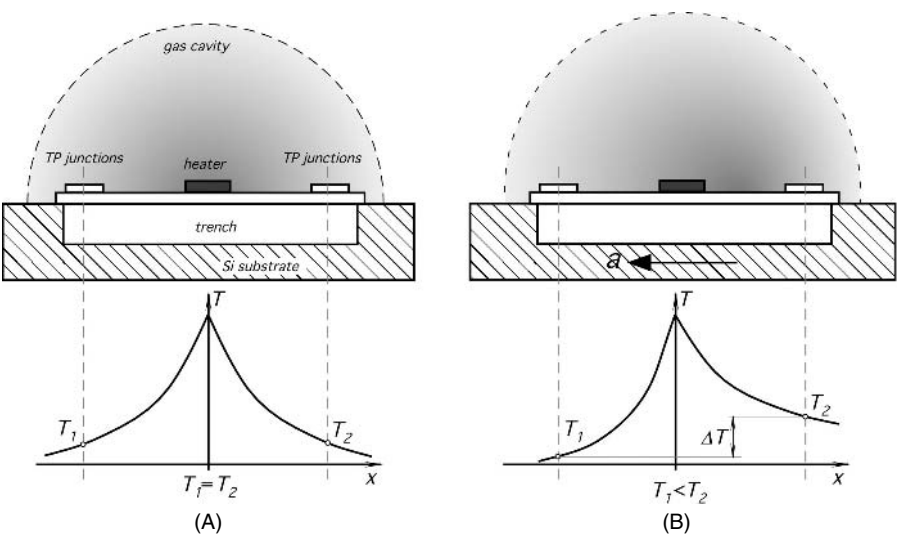


Fig. 8.8. (A) Cross-sectional view of the HGA sensor along the x axis. Heated gas is symmetrical around the heater. (B) Acceleration causes heated-gas shift to the right, resulting in a temperature gradient.

thermopile junctions (TPs) for sensing a temperature gradient along a single axis. Gas is heated so that it is hottest near the heater and rapidly cools down toward the left and right temperature sensors (thermopile junctions). When no force acts on gas, the temperature has a symmetrical conelike distribution around the heater, where temperatures T_1 at the left TP is equal to temperature T_2 of the right TP. Acceleration in any direction will disturb the temperature profile, due to a convection heat transfer, causing it to be asymmetrical. Figure 8.8B shows acceleration a in the direction of the arrow. Under the acceleration force, warm gaseous molecules shift toward the right TP and transfer a portion of their thermal energy to it. The temperature, and hence voltage, output of the opposite TP junctions will then be different, so $T_1 < T_2$. The differential temperature ΔT, and thus voltage, at the thermopile outputs becomes directly proportional to the acceleration. There are two identical acceleration signal paths on the device: one to measure acceleration along the x axis and one to measure acceleration along the y axis.

The HGA is capable of measuring accelerations with a full-scale range from below $\pm 1.0g$ to above $\pm 100g$. It can measure both the dynamic acceleration (e.g., vibration) and static acceleration (e.g., gravity). The analog output voltages from the chip are available in absolute and ratiometric modes. The absolute output voltage is independent of the supply voltage, and the ratiometric output voltage is proportional to the supply voltage. The typical noise floor is below 1 mg/Hz, allowing sub-milli-g signals to be measured at very low frequencies. The frequency response, or the capability to measure fast changes in acceleration, is defined by design. A typical -3-dB rolloff occurs at above 30 Hz, but it is expandable with a compensation to over 160 Hz.

It should be noted that for the HGA sensor, the output sensitivity changes with ambient temperature. The sensitivity change is shown in Fig. 8.9. To compensate for the change, an imbedded temperature sensor (a resistive temperature detector or silicon junction may be provided on the chip).

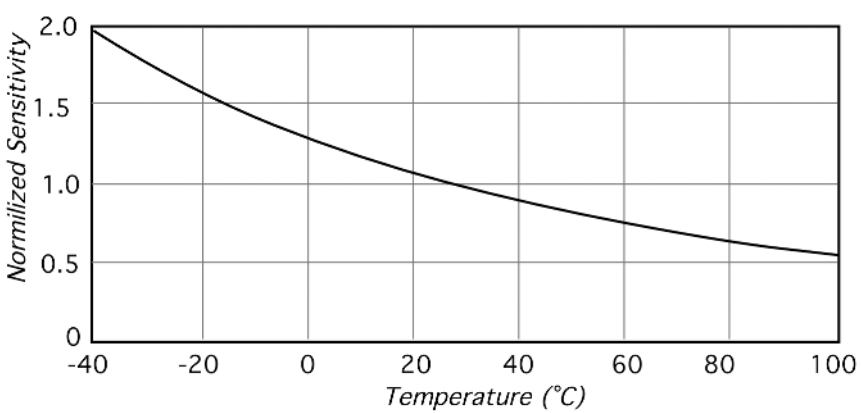


Fig. 8.9. Thermal accelerometer (HGA) sensitivity to ambient temperature.

8.6 Gyroscopes

Next to a magnetic compass, a gyroscope is probably the most common navigation sensor. In many cases, where a geomagnetic field is either absent (in space) or is altered by the presence of some disturbances, a gyroscope is an indispensable sensor for defining the position of a vehicle. A gyroscope, or a *gyro* for short, is a "keeper of direction," like a pendulum in a clock is a "keeper of time." A gyro operation is based on the fundamental principle of the conservation of angular momentum: *In any system of particles, the total angular momentum of the system relative to any point fixed in space remains constant, provided no external forces act on the system.*

8.6.1 Rotor Gyroscope

A mechanical gyro is comprised of a massive disk free to rotate about a spin axis (Fig. 8.10) which itself is confined within a framework that is free to rotate about one or two axes. Hence, depending on the number of rotating axes, gyros can be either of a single-, or two-degree-of-freedom type. The two qualities of a gyro account for it usefulness are as follows: (1) the spin axis of a free gyroscope will remain fixed with respect to space, provided there are no external forces to act upon it and (2) a gyro can be made to deliver a torque (or output signal) which is proportional to the angular velocity about an axis perpendicular to the spin axis.

When the wheel (rotor) freely rotates, it tends to preserve its axial position. If the gyro platform rotates around the input axis, the gyro will develop a torque around a perpendicular (output) axis, thus turning its spin axis around the output axis. This phenomenon is called the *precession* of a gyro. It can be explained by Newton's law of motion for rotation: *The time rate of change of angular momentum about any given axis is equal to the torque applied about the given axis*. That is to say, when a torque T is applied about the input axis, and the speed ω of the wheel is held constant, the

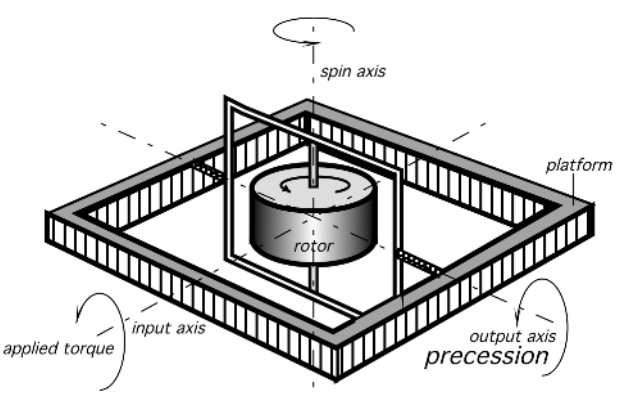


Fig. 8.10. Mechanical gyroscope with a single degree of freedom.

angular momentum of the rotor may be changed only by rotating the projection of the spin axis with respect to the input axis; that is, the rate of rotation of the spin axis about the output axis is proportional to the applied torque

$$T = I\omega\Omega, \tag{8.14}$$

where Ω is the angular velocity about the output axis and I is the inertia of a gyro wheel about the spin axis. To determine the direction of precession, the following rule can be used: *Precession is always in such a direction as to align the direction of rotation of the wheel with the direction of rotation of the applied torque*.

The accuracy of mechanical gyros heavily depends on the effects which may cause additional unwanted torques and cause drifts. The sources of these are friction, imbalanced rotor, magnetic effects, and so forth. One method which is widely used to minimize rotor friction is to eliminate the suspension entirely by floating the rotor and the driving motor in a viscous, high- density liquid, such as one of the fluorocarbons. This method requires close temperature control of the liquid and also may suffer from aging effects. The other method of friction reduction is to use the so-called gas bearings, where the shaft of the rotor is supported by high-pressure helium, hydrogen, or air. An even better solution is to support the rotor in vacuum by an electric field (electrostatic gyro). A magnetic gyro consists of a rotor supported by a magnetic field. In that case, the system is cryogenically cooled to temperatures where the rotor becomes superconductive. Then, an external magnetic field produces enough counterfield inside the rotor that the rotor floats in a vacuum. These magnetic gyroscopes also are called cryogenic.

8.6.2 Monolithic Silicon Gyroscopes

Although a spinning-rotor gyroscope was the only practical choice for many years, its operating principle really does not lend itself to the design of a small monolithic sensor that is required by many modern applications. Conventional spinning rotor gyroscopes contain parts such as gimbals, support bearings, motors, and rotors which need accurate machining and assembly; these aspects of construction prohibit conventional mechanical gyroscopes from ever becoming a low-cost device. Wear on the motors and bearings during operation means that the gyroscope will only meet the performance specifications for a set number of running hours. Other methods for sensing direction and velocity of motion have been developed. Often, a global positioning system (GPS) would be the ideal choice. Yet, frequently it just cannot be employed in space, under water, or whenever the size and cost are of paramount importance. Use of MEMS micromachine technology allows the design of a miniature gyroscope where the rotating disk is replaced with a vibrating element. The design takes advantage of the techniques developed in the electronic industry and is highly suited to high-volume manufacture. In addition, the vibrating gyro is much more robust and can withstand the environments typical of many military and aerospace applications.

All vibrating gyroscopes rely on the phenomenon of the Coriolis acceleration. The Coriolis effect is an inertial force described by the nineteenth-century French engineer–mathematician Gustave-Gaspard Coriolis in 1835. Coriolis showed that if

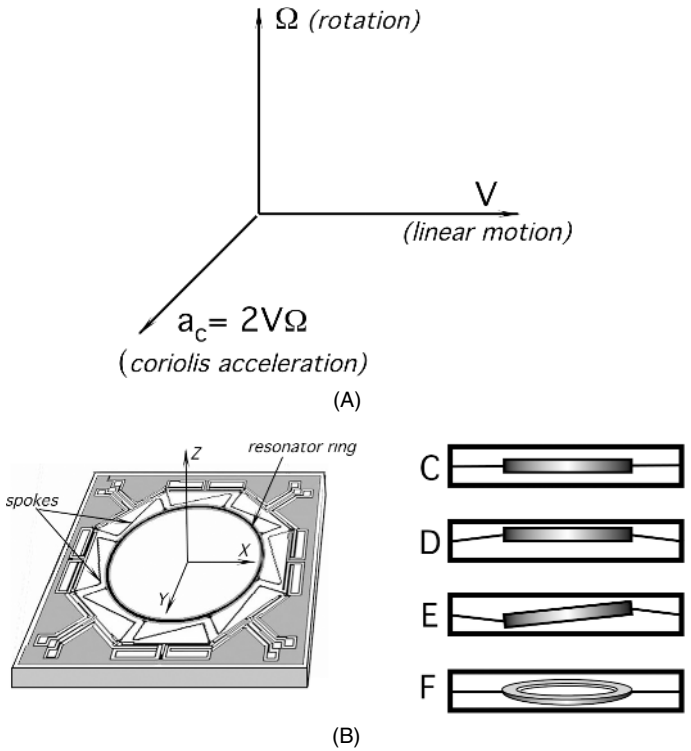


Fig. 8.11. (A) Coriolis acceleration; (B) Vibrating-ring micromachined structure; (C–F) effects of acceleration on the vibrating modes of the ring.

the ordinary Newtonian laws of motion of bodies are to be used in a rotating frame of reference, an inertial force, acting to the right of the direction of body motion for counterclockwise rotation of the reference frame or to the left for clockwise rotation, must be included in the equations of motion. The Coriolis acceleration of a body appears whenever that body moves linearly in a frame of reference which is rotating about an axis perpendicular to that of the linear motion. The resulting acceleration, which is directly proportional to the rate of turn, occurs in the third axis, which is perpendicular to the plane containing the other two axis (Fig. 8.11A). In a micromachined gyro, the rotation is replaced by vibration and the resulting acceleration can be detected and related to the rate of motion. Instead of a mass following a circular trajectory as for the conventional spinning—rotor gyroscope, the mass can be suspended and made to move linearly in simple harmonic motion.

There are several practical ways to build a vibrating gyro; however, all of them can be divided into three principle groups [6]:

1. Simple oscillators (mass on a string, beams)
2. Balanced oscillators (tuning forks)
3. Shell resonators (wine glass, cylinder, ring)

All three categories have been implemented in the actual designs.

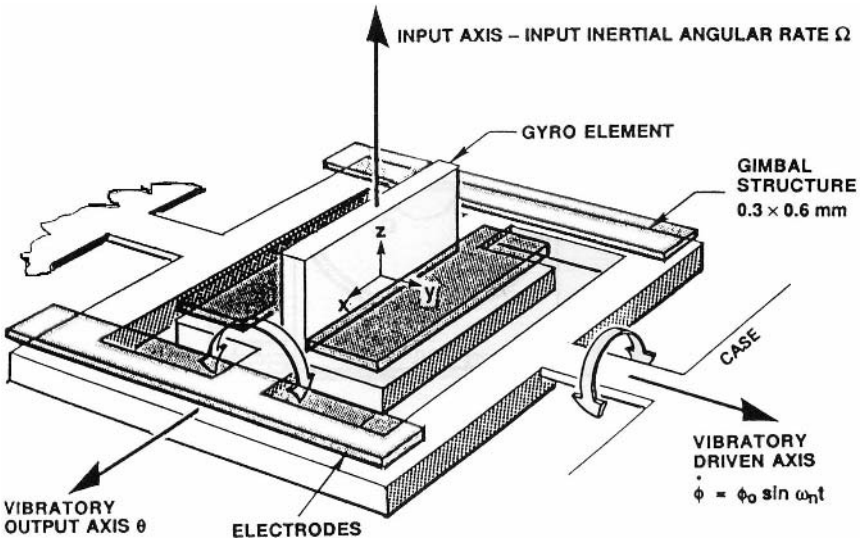


Fig. 8.12. Vibratory rate gyro concept. (From Ref. [7].)

One of the first such devices was a two-gimbal structure supported by torsional flexures (Fig. 8.12). It is undercut and free to move in the active area. In operation, the outer gimbal, or "motor", is driven at a constant amplitude by electrostatic torquing using electrodes placed in close proximity. This oscillatory motion is transferred to the inner gimbal along the stiff axis of the inner flexures, setting up an oscillating momentum vector with the inertial element. In the presence of an angular rotational rate normal to the plane of the device, the Coriolis force will cause the inner gimbal to oscillate about its weak axis with a frequency equal to the drive frequency and with an amplitude proportional to the inertial input rate. Maximum resolution is obtained when the outer gimbal is driven at a resonant frequency of the inner gimbal. The readout of the output motion is accomplished by setting the differential change in capacitance between the inner gimbal and a pair of electrodes. When operated in an open loop, the angular displacement of the inner gimbal about the output axis is proportional to the input rate; that is, the output angle Θ is proportional to an inertia ration term, the drive angle, ϕ_0, the mechanical Q, and the input rate Ω. It is inversely proportional to the drive frequency ω_n:

$$\Theta = \left[\frac{\mathbf{I}_x + \mathbf{I}_y - \mathbf{I}_z}{\mathbf{I}_x}\right] \frac{\phi_0 \Omega Q}{\omega_n}. \tag{8.15}$$

In a practical application, the device is operated closed loop and the inner gimbal is rebalanced to null in phase and in quadrature. A detailed description of the gyroscope may be found elsewhere [7].

A more recent design that also belongs to the third category was developed by British Aerospace Systems and Equipment along with its partner Sumitomo Precision

Products Company Ltd. [8]. The design is based on a ring resonator that is micromachined in silicon. Silicon has remarkable mechanical properties (see Section 18.1.1 of Chapter 18 for details); specifically, in its crystalline state, silicon has a fracture limit of 7 GPa, which is higher than the majority of steels. Coupled with this is a low density of 2330 kg/m^3, resulting in a very robust material under its own weight. The gyro resonator is etched out of the crystalline material. This ensures that the properties of the resonator are stable over its lifetime and environment. The planar vibrating-ring structure has all of the vibration energy in one plane. As such, under angular rate, there is no coupling of vibration from one crystal plane to another, so that the vibrating parameters are very stable over temperature.

In order for the resonator to function correctly, it must be supported in a way that allows it to vibrate as freely as possible. The sensing element is shown in Fig. 8.11B. The resonator comprises a 6-mm silicon ring, supported by eight radially compliant spokes, which are anchored to a 10×10-mm support frame. Current-carrying conductors are deposited and patterned onto the top surface only, and pads for wire bonding are located on the outer support frame. The chip is anodically bonded to a supporting glass structure which is thermally matched to the silicon. There are eight identical conducting loops, each of which follows the pattern: bond pad→along length of support leg→around 1/8 segment of ring→along length of next support leg→bond pad. Each leg thus contains two conductors, one each from adjacent loops, in addition to a third conductor, which lies between them, to minimize capacitive coupling. The silicon substrate is also connected in order to provide a ground plane. The resonator may be excited into vibration by any suitable transducers. These may function by means of optical, thermal expansion, piezoelectric, electrostatic or electromagnetic effects, for example. The excitation may be applied to the support structure which carries the resonator, or directly to the resonator itself. The fundamental vibration mode is at 14.5 kHz. Figures 8.11C–8.11F show the effects of linear and angular acceleration on the resonator. Figure 8.11C shows a side view of the resonator under conditions of no acceleration, Fig. 8.11D shows the effect of z-axis linear acceleration, Fig. 8.11E shows the effect of angular acceleration about the x axis, and Fig. 8.11F shows the effect of angular acceleration about the y axis. Because the ring position changes with respect to the frame, what is required is a combination of displacement pickup transducers to detect a particular movement of the resonator. Resonator vibration may, for example, be sensed by transducers working electromagnetically, capacitively, optically, piezoelectrically, or by means of strain gauges. In this particular design, a magnetic pickup is employed by pattern conductive loops along with a magnetic field which is perpendicular to the plane of the ring. The magnetic field is provided by samarium cobalt and the entire structure is housed in a standard hermetic metal integrated circuit can package.

8.6.3 Optical Gyroscopes

Modern development of sensors for guidance and control applications is based on employing the so-called Sagnac effect, which is illustrated in Fig. 8.13 [9]. Two beams of light generated by a laser propagate in opposite directions within an optical

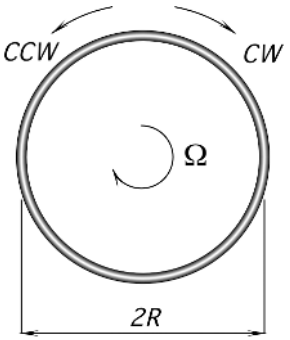


Fig. 8.13. Sagnac effect.

ring having refractive index n and radius R. One beam goes in a clockwise (CW) direction, and the other goes in a counterclockwise (CCW) direction. The amount of time it takes light to travel within the ring is $\Delta t = 2\pi R/nc$, where c is the speed of light. Now, let us assume that the ring rotates with angular rate Ω in the clockwise direction. In that case, light will travel different paths at two directions. The CW beam will travel $l_{\text{cw}} = 2\pi R + \Omega R\Delta t$, and the CCW beam will travel $l_{\text{ccw}} = 2\pi R - \Omega R\Delta t$. Hence, the difference between the paths is

$$\Delta l = \frac{4\pi\Omega R^2}{nc}. \tag{8.16}$$

Therefore, to accurately measure Ω, a technique must be developed to determine Δl. There are three basic methods known for the path detection: (1) optical resonators, (2) open-loop interferometers, and (3) closed-loop interferometers.

For the ring laser gyro, measurements of Δl are made by taking advantages of the lasing characteristics of an optical cavity (i.e., of its ability to produce coherent light). For lasing to occur in a closed optical cavity, there must be an integral number of wavelengths about the complete ring. The light beams, which do not satisfy this condition, interfere with themselves as they subsequently travel the optical path. In order to compensate for a change in the perimeter due to rotation, the wavelength λ and frequency ν of the light must change:

$$-\frac{d\nu}{\nu} = \frac{d\lambda}{\lambda} = \frac{dl}{l}. \tag{8.17}$$

Equation (8.17) is a fundamental equation relating frequency, wavelength, and perimeter change in the ring laser. If the ring laser rotates at a rate Ω, then Eq. (8.16) indicates that light waves stretch in one direction and compress in the other direction to meet the criteria for the lasing of an integral number of wavelengths about the ring. This, in turn, results in a net frequency difference between the light beams. If the two beams are bit together (mixed), the resulting signal has frequency is

$$F = \frac{4A\Omega}{\lambda nl}, \tag{8.18}$$

where A is the area enclosed by the ring.

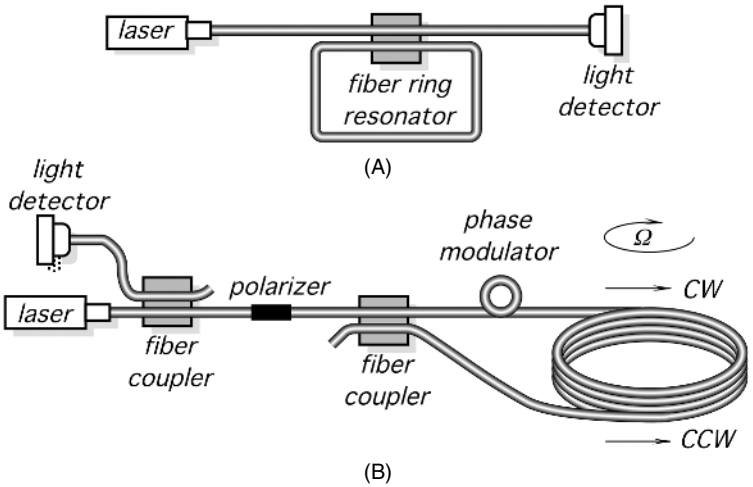


Fig. 8.14. (A) Fiber-optic ring resonator; (B) fiber-optic analog coil gyro. (Adapted from Ref. [9].)

In practice, optic gyros are designed with either a fiber ring resonator, or a fiber coil where the ring has many turns of the optical fiber [10]. The optic ring resonator is shown in Fig. 8.14A. It consists of a fiber loop formed by a fiber beam splitter that has a very low cross-coupling ratio. When the incoming beam is at the resonant frequency of the fiber ring, the light couples into the fiber cavity and the intensity in the exiting light drops. The coil fiber gyro (Fig. 8.14B) contains a light source and the detector coupled to the fiber. The light polarizer is positioned between the detector and the second coupler to ensure that both counterpropagating beams traverse the same path in the fiber-optic coil [11]. The two beams mix and impinge onto the detector, which monitors the cosinusoidal intensity changes caused by rotationally induced phase changes between the beams. This type of optical gyro provides a relatively low-cost, small-size, rotation-sensitive sensor with a dynamic range up to 10,000. Applications include yaw and pitch measurements, attitude stabilization, and gyrocompassing. A major advantage of optical gyros is their ability to operate under hostile environments that would be difficult, if not impossible, for the mechanical gyros.

8.7 Piezoelectric Cables

A piezoelectric effect is employed in a vibration sensor built with a mineral-insulated cable. Such a cable generates an electric signal in its internal conductor when the outer surface of the cable is compressed. The piezoelectric Vibracoax™ cables[5] have been used in various experiments to monitor the vibration in compressor blades

[5] Philips Electronic Instruments, Norcross, GA.

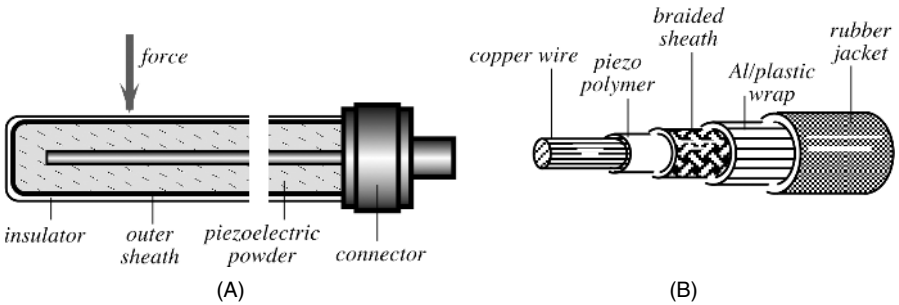


Fig. 8.15. Piezoelectric cable sensors: (A) construction of Vibracoax; (B) polymer film as a voltage generating component. (Adapted from [13].)

in turboshaft aircraft engines. Other applications include the detection of insects in silos and automobile traffic analysis. In these applications, the cables are buried in the highway pavement, positioned perpendicular to the traffic. When properly installed, they last for at least 5 years [12]. The sensors are designed to be sensitive primarily to vertical forces. A piezoelectric cable consists of a solid insulated copper sheath having a 3-mm outer diameter, piezoelectric ceramic powder, and an inner copper core (Fig. 8.15A). The powder is tightly compressed between the outer sheath and the core. Usually, the cable is welded at one end and connected to a 50Ω extension cable at the other end.

Another method of fabrication of the piezoelectric cables is to use a polyvinylidene fluoride (PVDF) polymer film as a component in the cable insulation (Fig. 8.15B). The PVDF can be made piezoelectric, thus giving the cable sensing properties. When a mechanical force is applied to the cable, the piezoelectric film is stressed, which results in the development of electric charges of the opposite polarities on it surfaces. The inner copper wire and the braided sheath serve as charge pickup electrodes.

For the cable to possess piezoelectric properties, its sensing component (the ceramic powder or polymer film) must be poled during the manufacturing process; that is, the cable is warmed up to near the Curie temperature, and subjected to high voltage to orient ceramic dipoles in the powder or polymer dipoles in the film, then cooled down while the high voltage is maintained. When the cable sensor is installed in the pavement (Fig. 8.16), its response should be calibrated, because the shape of the signal and its amplitude depend not only on the properties of the cable but also on the type of the pavement and subgrade. The electrical output is proportional to the stress imparted to the cable. The long, thin piezoelectric insulating layer provides a relatively low output impedance (600 pF/m), unusual for a piezoelectric device. The dynamic range of the cable is substantial ($>$200 dB), sensing distant, small-amplitude vibrations caused by rain or hail, yet responding linearly to the impacts of heavy trucks. The cables have withstood pressures of 100 MPa. The typical operating temperature range is -40°C to $+125$°C. Table 8.2 lists typical properties for piezo cable.

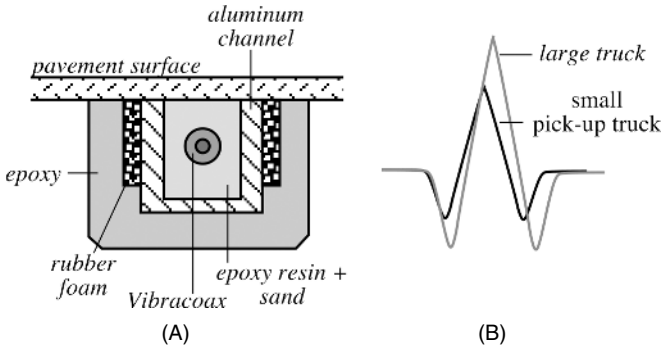


Fig. 8.16. Application of the piezoelectric cables in highway monitoring: (A) sensor installation in the pavement; (B) shape of electrical response.

Table 8.2. Typical Properties of a Piezoelectric Cable

Parameter	Units	Value
Capacitance at 1 kHz	pF/m	600
Tensile strength	MPa	60
Young's modulus	GPa	2.3
Density	kg/m^3	1890
Acoustic impedance	MRayl (10^6/sm^2)	4.0
Relative permittivity	1 kHz	9
tan δ_e	1 kHz	0.017
Hydrostatic piezocoefficient	pC/N	15
Longitudinal piezocoefficient	V m/N	250×10^{-3}
Hydrostatic piezocoefficient	V m/N	150×10^{-3}
Electromechanical coupling	%	20
Energy output	mJ/strain (%)	10
Voltage output	kV/strain (%)	5

Source: Ref. [14].

References

1. Articolo, G. A. Shock impulse response of a force balance servo-accelerometer. In: *Sensors Expo West Proceedings*. Helmers Publishing, 1989.
2. Sensor signal conditioning: an IC designer's perspective. *Sensors Magazine*, 23–30, 1991.
3. Allen, H., Terry, S., and De Bruin, D. Accelerometer system with self-testable features. *Sensors Actuators* 20, 153–161, 1989.
4. Suminto, J. T. A simple, high performance piezoresistive accelerometer. In: *Transducers'91. 1991 International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*., IEEE, New York, 1991, pp: 104–107.

5. Haritsuka, R., van Duyn, D.S., Otaredian, T., and de Vries, P. A novel accelerometer based on a silicon thermopile. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*. IEEE, New York, 1991, pp: 420–423.
6. Fox, C.H.J. and Hardie, D.S.W. Vibratory gyroscopic sensors. *Symposium Gyro Technology* (DGON), 1984.
7. Boxenhom, B.B., Dew. B., and Greiff, P. The micromechanical inertial guidance system and its applications. In: *14th Biennial Guidance Test Symposium*, 6588th Test Group, Holloman AFB, New Mexico, Oct. 3–5, 1989.
8. Varnham, M.P., Hodgins, D., Norris, T.S., and Thomas, H.D. Vibrating planar gyro. U.S. patent 5,226,321, 1993.
9. Udd, E. Fiber optic sensors based on the Sagnac interferometer and passive ring resonator. In: *Fiber Optic Sensors*. E. Udd, ed. John Wiley & Sons, New York, 1991, pp. 233–269.
10. Ezekiel, S. and Arditty, H, J., eds. *Fiber-Optic Rotation Sensors*. Springer Series in Optical Sciences. Vol. 32, Springer-Verlag, New York, 1982.
11. Fredericks, R. J., and Ulrich, R. Phase error bounds of fiber gyro with imperfect polarizer/depolarizer. *Electron. Lett.* 29, 330, 1984.
12. Bailleul, G. Vibracoax piezoelectric sensors for road traffic analysis. *Sensor Expo Proceedings*, Helmers Publishing, 1991.
13. Radice, P. F. Piezoelectric sensors and smart highways. In: *Sensors Expo Proceedings*. Helmers Publishing, 1991.
14. *Piezo Film Sensors Technical Manual*. Measurement Specialties, Inc., Fairfield, NJ., April 1999; available at www.msiusa.com.

9

Force, Strain, and Tactile Sensors

Whereas kinematics studies positions of objects and their motions, dynamics answers the question "What causes the motion?" Classical mechanics deal with moving objects whose velocities are substantially smaller than the speed of light. Moving particles, such as atoms and electrons, are the subject of quantum mechanics and the theory of relativity. A typical problem of classical mechanics is the question: "What is motion of an object, which initially had a given mass, charge, dipole moment, position, and so forth and was subjected to external objects having known mass, charge, velocity, and so forth?" That is, to say, classical mechanics deals with interactions of macroobjects. In a general form, this problem was solved by Sir Isaac Newton (1642–1727), who claimed he was born in the year when Galileo died.[1] He brilliantly developed ideas of Galileo and other great mechanics. Newton stated his first law as: *Every body persists in its state of rest or of uniform motion in a straight line unless it is compelled to change that state by forces impressed on it.* Sometimes, this is called the law of inertia. Another way to state the first law is to say: "If no net force acts on a body, its acceleration **a** is zero."

When force is applied to a free body (not anchored to another body), it gives the body an acceleration in the direction of force. Thus, we can define force as a vector value. Newton had found that acceleration is proportional to the acting force **F** and inversely proportional to the property of a body called the *mass m* which is a scalar value:

$$\mathbf{a} = \frac{\mathbf{F}}{m}. \tag{9.1}$$

This equation is known as *Newton's second law*; the name was given by the great Swiss mathematician and physicist Leonhard Euler in 1752, 65 years after the publication of Newton's *Principia* [1]. The first law is contained in the second law as a special case: When net acting force $\mathbf{F} = 0$, acceleration $\mathbf{a} = 0$.

Newton's second law allows us to establish the mechanical units. In SI terms, mass (kg), length (m), and time (s) are the *base* units (see Table 1.7 of Chapter 1). Force and acceleration are *derivative* units. The force unit is the force which will accelerate 1 kg mass to acceleration 1 m/s^2. This unit is called a *newton*.

[1] In reality, Newton was born on January 4, 1643.

Table 9.1. Mechanical Units

System of Units	Force	Mass	Acceleration
SI	Newton (N)	**kilogram** (kg)	m/s^2
British	**Pound** (lb)	Slug	ft/s^2

Note: Boldface indicates base units.

In the British and U.S. Customary systems of units, however, force (lb), length (ft), and time (s) are selected as the base units. The mass unit is defined as the mass which is accelerated at 1 ft/s^2 when it is subjected to force of 1 lb. The British unit of mass is *slug*. The mechanical units are as shown in Table 9.1.

Newton's third law establishes the principle of a mutual interaction between two bodies: *To every action there is always opposed an equal reaction; or, the mutual actions of two bodies upon each other are always equal, and directed to contrary parts*.

In engineering measurements, it is often necessary to know the density of a medium, which is amount of matter per unit volume. Density is defined through mass m and volume V as

$$\rho = \frac{m}{V}. \tag{9.2}$$

The unit of density is kg/m^3 or lb/ft^3 (British system). Densities of some materials are given in Table A.12 of the Appendix.

The SI unit of force is one of the fundamental quantities of physics. The measurement of force is required in mechanical and civil engineering, for weighing objects, designing prosthesis, and so forth. Whenever pressure is measured, it requires the measurement of force. It could be said that force is measured when dealing with solids, whereas pressure is measured when dealing with fluids (i.e., liquids or gases); that is, force is considered when action is applied to a spot, and pressure is measured when force is distributed over a relatively large area.

Force sensors can be divided into two classes: quantitative and qualitative. A quantitive sensor actually measures the force and represents its value in terms of an electrical signal. Examples of these sensors are strain gauges and load cells. The qualitative sensors are the *threshold* devices which are not concerned with a good fidelity of representation of the force value. Their function is merely to indicate whether a sufficiently strong force is applied; that is, the output signal indicates when the force's magnitude exceeds a predetermined threshold level. An example of these detectors is a computer keyboard, on which a key makes a contact only when it is pressed sufficiently hard. The qualitative force sensors are frequently used for the detection of motion and position, as described in Chapter 7. A pressure-sensitive floor mat and a piezoelectric cable are examples of the qualitative pressure sensors.

The various methods of sensing force can be categorized as follows [2]:

1. By balancing the unknown force against the gravitational force of a standard mass
2. By measuring the acceleration of a known mass to which the force is applied

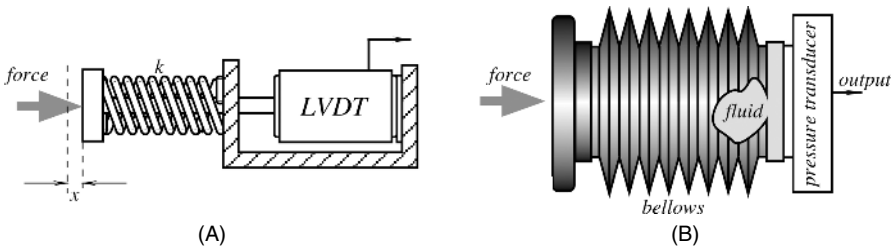


Fig. 9.1. (A) Spring-loaded force sensor with LVDT; (B) force sensor with a pressure transducer.

3. By balancing the force against an electromagnetically developed force
4. By converting the force to a fluid pressure and measuring that pressure
5. By measuring the strain produced in a elastic member by the unknown force

In modern sensors, the most commonly used method is method 5; methods 3 and 4 are used occasionally.

In most sensors, force is not directly converted into an electric signal. Some intermediate steps are usually required. Thus, many force sensors are *complex* sensors. For instance, a force sensor can be fabricated by combining a force-to-displacement converter and a position (displacement) sensor. The former may be a simple coil spring, whose compression displacement x can be defined through the spring coefficient k and compressing force F as

$$x = kF. \tag{9.3}$$

The sensor shown in Fig. 9.1A is composed of a spring and linear variable differential transformer (LVDT) displacement sensor (Section 7.4 of Chapter 7). Within the linear range of the spring, the LVDT sensor produces a voltage which is proportional to the applied force. A similar sensor can be constructed with other types of spring and pressure sensor, such as the one shown in Fig. 9.1B. The pressure sensor is combined with a fluid-filled bellows which is subjected to force. The fluid-filled bellows functions as a force-to-pressure converter by distributing a localized force at its input over the sensing membrane of a pressure transducer.

9.1 Strain Gauges

A strain gauge is a resistive elastic sensor whose resistance is a function of applied strain (unit deformation). Because all materials resist deformation, some force must be applied to cause deformation. Hence, resistance can be related to applied force. That relationship is generally called the *piezoresistive* effect (see Section 3.5.3 of Chapter 3) and is expressed through the gauge factor S_e of the conductor [Eq. (3.63) of Chapter 3]:

$$\frac{dR}{R} = S_e e, \tag{9.4}$$

For many materials, $S_e \approx 2$ with the exception of platinum, for which $S_e \approx 6$ [3].

Table 9.2. Characteristics of Some Resistance Strain Gauges

Material	Gauge factor (S_e)	Resistance, Ω	Temperature coefficient of resistance ($°C^{-1} \times 10^{-6}$)	Notes
57% Cu–43%Ni	2.0	100	10.8	S_e is constant over a wide range of strain; for use under 260°C
Platinum alloys	4.0–6.0	50	2,160	For high-temperature use
Silicon	−100 to +150	200	90,000	High sensitivity, good for large strain measurements

Source: Ref. [4].

For small variations in resistance not exceeding 2% (which is usually the case), the resistance of a metallic wire is

$$R = R_0(1 + x), \tag{9.5}$$

where R_0 is the resistance with no stress applied, and $x = S_e e$. For the semiconductive materials, the relationship depends on the doping concentration (Fig. 18.2A of Chapter 18). Resistance decreases with compression and increases with tension. Characteristics of some resistance strain gauges are given in Table 9.2.

A wire strain gauge is composed of a resistor bonded with an elastic carrier (backing). The backing, in turn, is applied to the object for which stress or force should be measured. Obviously, that strain from the object must be reliably coupled to the gauge wire, whereas the wire must be electrically isolated from the object. The coefficient of thermal expansion of the backing should be matched to that of the wire. Many metals can be used to fabricate strain gauges. The most common materials are alloys *constantan, nichrome, advance,* and *karma*. Typical resistances vary from 100 Ω to several thousand ohms. To possess good sensitivity, the sensor should have long longitudinal and short transverse segments (Fig. 9.2), so that transverse sensitivity is no more than a couple of percent of the longitudinal. The gauges may be arranged in many ways to measure strains in different axes. Typically, they are connected into Wheatstone bridge circuits (Section 5.7 of Chapter 5). It should be noted that semiconductive strain gauges are quite sensitive to temperature variations. Therefore, interface circuits or the gauges must contain temperature-compensating networks.

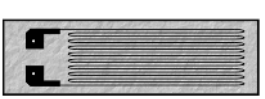


Fig. 9.2. Wire strain gauge bonded on elastic backing.

9.2 Tactile Sensors

Tactile sensors are a special class of force or pressure transducers, which are characterized by small thickness. This makes the sensors useful in applications where force or pressure can be developed between two surfaces being in close proximity to one another. Examples include robotics, where tactile sensors can be positioned on the "fingertips" of a mechanical actuator to provide a feedback upon developing a contact with an object—very much like tactile sensors work in human skin. They can be used to fabricate "touch screen" displays, keyboards, and other devices where a physical contact has to be sensed. A very broad area of applications is in the biomedical field, where tactile sensors can be used in dentistry for the crown or bridge occlusion investigation and in studies of forces developed by a human foot during locomotion. They can be installed in artificial knees for the balancing of the prosthesis operation and so on. In mechanical and civil engineering, the sensors can be used to study forces developed by fastening devices.

Several methods can be used to fabricate tactile sensors. Some of them require the formation of a thin layer of a material which is responsive to strain. A simple tactile sensor producing an "on–off" output can be formed with two leaves of foil and a spacer (Fig. 9.3). The spacer has round (or any other suitable shape) holes. One leaf is grounded and the other is connected to a pull-up resistor. A multiplexer can be used if more than one sensing area is required. When an external force is applied to the upper conductor over the opening in the spacer layer, the conductor flexes, and upon reaching the lower conductor, it makes an electric contact, grounded by that the pull-up resistor. The output signal becomes zero, indicating the applied force. The upper and lower conducting leaves can be fabricated by a silkscreen printing of conductive ink on the backing material, like Mylar® or polypropylene. Multiple sensing spots can be formed by printing rows and columns of a conductive ink. Touching of a particular

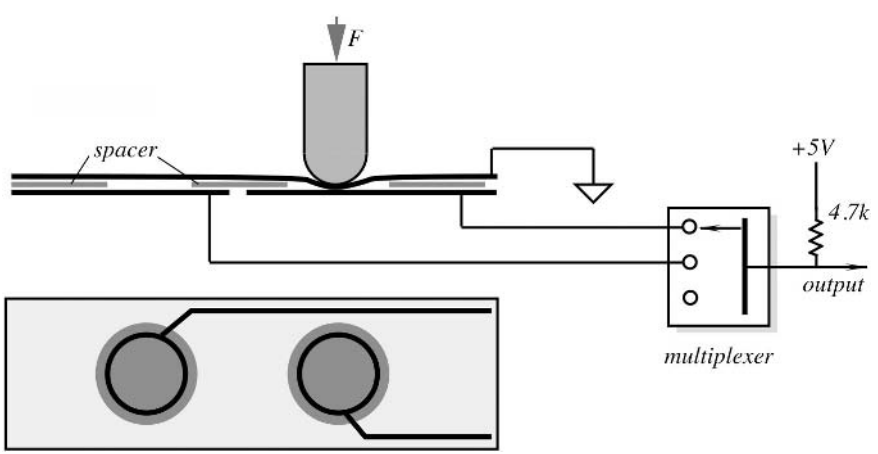


Fig. 9.3. Membrane switch as a tactile sensor.

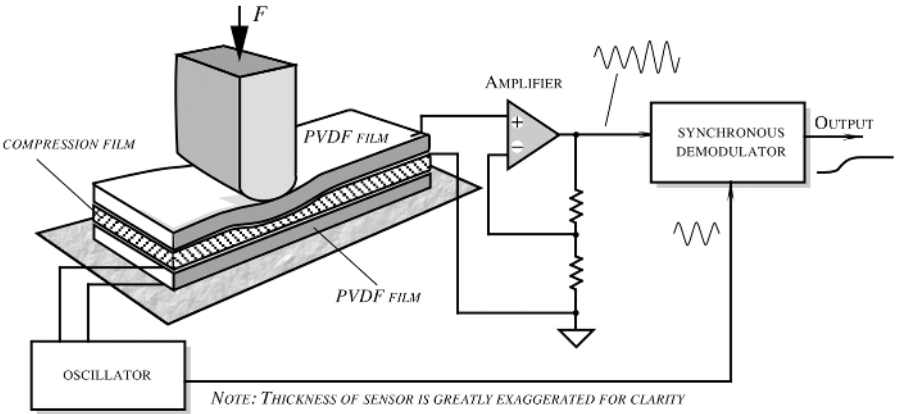


Fig. 9.4. Active piezoelectric tactile sensor.

area on a sensor will cause the corresponding row and column to join, thus indicating force at a particular location.

Good tactile sensors can be designed with piezoelectric films, such as polyvinylidene fluoride (PVDF) used in active or passive modes. An active ultrasonic coupling touch sensor with the piezoelectric films is illustrated in Fig. 9.4, in which three films are laminated together (the sensors also have additional protective layers which are not shown in the figure). The upper and the bottom films are PVDF, and the center film is for the acoustic coupling between the other two. The softness of the center film determines the sensitivity and the operating range of the sensor. The bottom piezoelectric film is driven by an ac voltage from an oscillator. This excitation signal results in mechanical contractions of the film which are coupled to the compression film and, in turn, to the upper piezoelectric film, which acts as a receiver. Because piezoelectricity is a reversible phenomenon, the upper film produces alternating voltage upon being subjected to mechanical vibrations from the compression film. These oscillations are amplified and fed into a synchronous demodulator. The demodulator is sensitive to both the amplitude and the phase of the received signal. When compressing force **F** is applied to the upper film, mechanical coupling between the three-layer assembly changes. This affects the amplitude and the phase of the received signal. These changes are recognized by the demodulator and appear at its output as a variable voltage.

Within certain limits, the output signal linearly depends on the force. If 25-μm PVDF films are laminated with a 40-μm silicone rubber compression film, the thickness of the entire assembly (including protective layers) does not exceed 200 μm. The PVDF film electrodes may be fabricated with a cell-like pattern on either the transmitting or receiving side. This would allow us to use electronic multiplexing of the cells to achieve spatial recognition of applied stimuli. The sensor also can be used to measure small displacements. Its accuracy is better than ± 2 μm over a range of a few millimeters. The advantages of this sensor is in its simplicity and a dc response, (i.e., in the ability to recognize static forces).

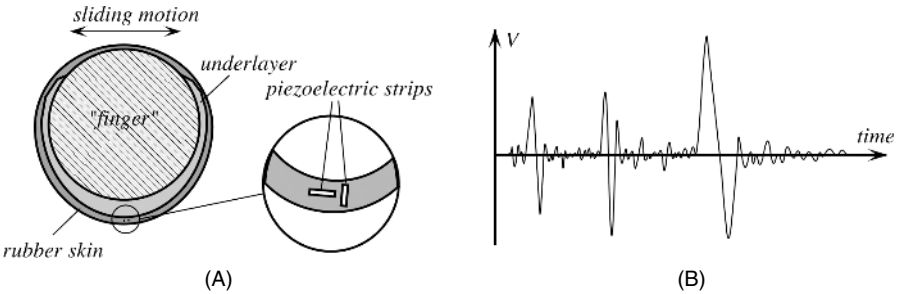


Fig. 9.5. Tactile sensor with a piezoelectric film for detecting sliding forces: (A) cross-sectional view; (B) typical response. (Adapted from Ref. [5].)

A piezoelectric tactile sensor can be fabricated with the PVDF film strips imbedded in a rubber skin (Fig. 9.5A). This sensor is passive; that is, its output signal is generated by the piezoelectric film without the need for an excitation signal. As a result, it produces a response proportional to the rate of stress, rather than to the stress magnitude. The design of this sensor is geared to robotic applications, where it is desirable to sense sliding motions causing fast vibrations. The piezoelectric sensor is directly interfaced with a rubber skin; thus, the electric signal produced by the strips reflect movements of the elastic rubber which results from the friction forces.

The sensor is built on a rigid structure (a robot's "finger") which has a foamy, compliant underlayer (1 mm thick), around which a silicon rubber "skin" is wrapped. It is also possible to use a fluid underlayer for better smooth-surface tracking. Because the sensing strips are located at some depth beneath the skin surface and because the piezoelectric film responds differently in different directions, the signal magnitude is not the same for movements in any direction. The sensor responds with a bipolar signal (Fig. 9.5B) to surface discontinuity or bumps as low as 50 μm high.

The following are few more examples of sensors that use PVDF and copolymer films [6]. Many tactile sensors are just sensitive conventional switches. However, the reliability of conventional contact switches is reduced due to contaminates like moisture and dust which foul the contact points. A piezoelectric film offers exceptional reliability, as it is a monolithic structure, not susceptible to this and other conventional switch failure modes. One of the most challenging of all switch applications is found in pinball machines. A pinball machine manufacturer uses a piezo film switch as a replacement for the momentary rollover-type switch. The switch is constructed from a laminated piezoelectric film on a spring steel beam, mounted as a cantilever to the end of a circuit board (Fig. 9.6A). The "digital" piezoelectric film switch is connected to a simple MOSFET circuit that consumes no power during the normally open state. In response to a direct contact force, the film beam momentarily triggers the MOSFET. This provides a momentary "high" state of the switch. The sensor does not exhibit the corrosion, pitting, or bounce that are normally associated with contact switches. It can survive in excess of 10 million cycles without failure. The simplicity of the design makes it effective in applications which include counter switches for assembly lines

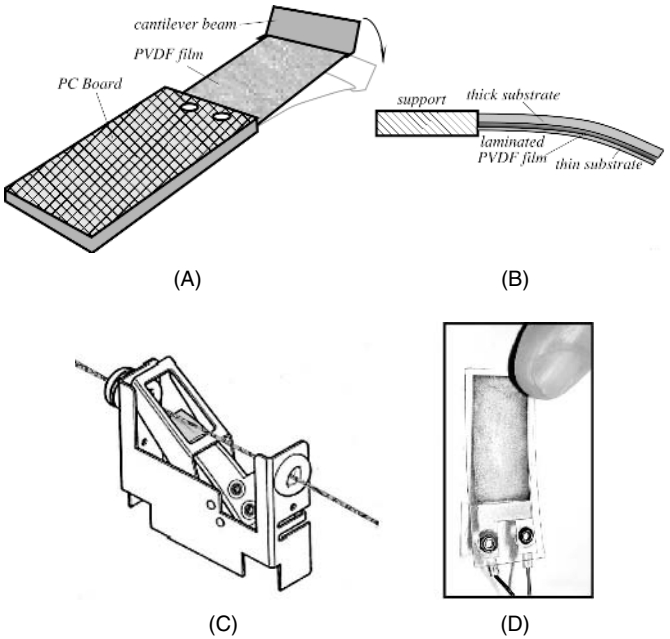


Fig. 9.6. (A) PVDF film switch for a pinball machine; (B) beam switch; (C) threadbreak sensor (from Ref. [6]); (D) PVDF tactile sensor.

and shaft rotation, switches for automated processes, impact detection for machine dispensed products, and so forth. The cantilever beam that carries the PVDF film can be modified to adjust switch sensitivity for high to low impact forces. Figure 9.6B shows the construction of the beam-type switch. The PVDF film element is laminated to a thicker substrate on one side and has a much thinner laminated substrate on the other. This moves the neutral axis of the structure out of the piezoelectric film element, resulting in a fully tensile strain in the film when deflected upward and a fully compressive strain when deflected in the opposite direction. Beam switches are used in shaft-rotation counters in natural gas meters and as gear-tooth counters in electric utility metering. The beam switch does not require an external power source, so the gas meter is safe from spark hazard. Other examples of applications for the beam switch include a baseball target that detects ball impact, a basketball game where a hoop-mounted piezoelectric film sensor counts good baskets, switches inside of an interactive soft doll to detect a kiss to the cheek or a tickle (the sensor is sewn into the fabric of the doll), coin sensors for vending and slot machines, and as a digital potentiometer for high reliability.

The popularity of electronics for musical instruments presents a special problem in drums and pianos. The very high dynamic range and frequency response requirements for drum triggers and piano keyboards are met by piezoelectric film impact elements. Laminates of piezo film are incorporated in the foot pedal switches for bass drums and in triggers for snares and tom-toms. Piezoelectric film impact switches are force

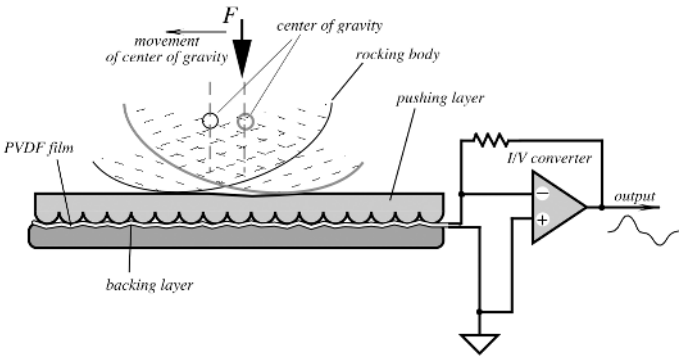


Fig. 9.7. Piezoelectric film respiration sensor.

sensitive, faithfully duplicating the effort of the drummer or pianist. In electronic pianos, the piezoelectric film switches respond with a dynamic range and time constant that is remarkably similar to a piano key stroke.

Textile plants require the continuous monitoring of often thousands of lines of thread for breakage. An undetected break event can require that a large volume of material be discarded, as the labor costs to recover the material exceed the manufacturing cost. Drop switches, where switch contact closure occurs when the thread breaks, are very unreliable. Lint fouls the contact points, resulting in no output signal. A piezoelectric film vibration sensor, mounted to a thin steel beam, monitors the acoustic signal caused by the abrasion of the thread running across the beam, analogous to a violin string (Fig. 9.6C). The absence of the vibration instantly triggers the machinery to stop.

Figure 9.7 shows a PVDF film tactile sensor for detecting the breathing rate of a sleeping child, where minute movements of the body resulted from respiration had to be monitored in order to detect cessation of breathing [7]. The sensor was placed under the mattress in a crib. The body of a normally breathing child slightly shifts with each inhale and exhale due to a moving diaphragm. This results in a displacement of the body's center of gravity which is detected by the sensor. The sensor consists of three layers, where the PVDF film is laminated between a backing material (e.g., silicone rubber) and a pushing layer. The pushing layer is fabricated of a plastic film (i.e., Mylar), whose side facing the PVDF film is preformed to have a corrugated surface. Under the variable force, the PVDF film is variably stressed by the groves of the pusher. This results in the generation by the film of electric charge. The charge flows out of the film through a current-to-voltage (I/V) converter which produces variable output voltage. The amplitude of that voltage within certain limits is proportional to the applied force.

Another type of tactile sensor is a piezoresistive sensor. It can be fabricated by using materials whose electrical resistance is a function of strain. The sensor incorporates a force-sensitive resistor (FSR) whose resistance varies with applied pressure [8]. Such materials are conductive elastomers or pressure-sensitive inks. A conductive

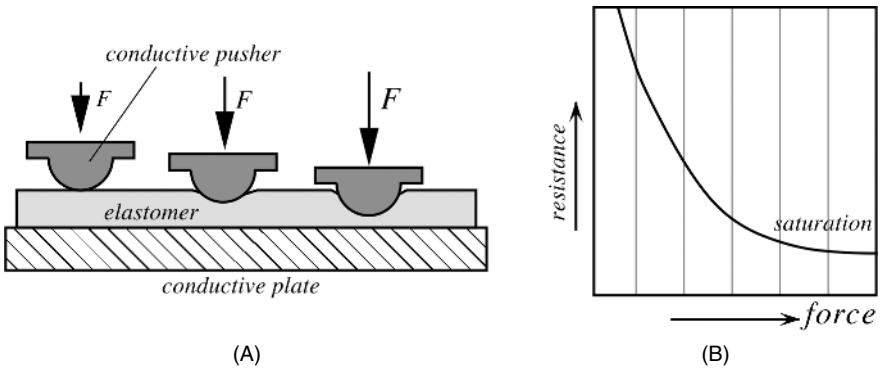


Fig. 9.8. FSR tactile sensor: (A) through-thickness application with an elastomer; (B) transfer function.

elastomer is fabricated of silicone rubber, polyurethane, and other compounds which are impregnated with conductive particles or fibers. For instance, conductive rubber can be fabricated by using carbon powder as an impregnating material. The operating principles of elastomeric tactile sensors are based either on varying the contact area when the elastomer is squeezed between two conductive plates (Fig. 9.8A) or in changing the thickness. When the external force varies, the contact area at the interface between the pusher and the elastomer changes, resulting in a reduction of electrical resistance. At a certain pressure, the contact area reaches its maximum and the transfer function (Fig. 9.8B) goes to saturation. For a resistive polymer Velostat^{TM} (from 3M), of thickness 70 μm and a specific resistance of 11 kΩ/cm^2, resistance for pressures over 16 kPa can be approximated by

$$R = \frac{51.93}{p^{1.47}} + 19. \tag{9.6}$$

It should be noted, however, that the resistance may noticeably drift when the polymer is subjected to prolonged pressure. Thus, the FSR sensors would be much more useful for qualitative rather than quantitative measurements.

A much thinner FSR tactile sensor can be fabricated with a semiconductive polymer whose resistance varies with pressure. A design of the sensor resembles a membrane switch (Fig. 9.9) [9]. Compared with a strain gauge, the FSR has a much wider dynamic range: typically three decades of resistanceЇchange over a 0–3-kg force range and much lower accuracy (typically ±10%). However, in many applications, where an accurate force measurement is not required, the very low cost of the sensor makes it an attractive alternative. A typical thickness of a FSR polymer sensor is in the range of 0.25 mm (0.010 in.), but much thinner sheets are also available.

Miniature tactile sensors are especially in high demand in robotics, where good spatial resolution, high sensitivity, and a wide dynamic range are required. A plastic deformation in silicon can be used for the fabrication of a threshold tactile sensor with a mechanical hysteresis. In one design [10], the expansion of trapped gas in a

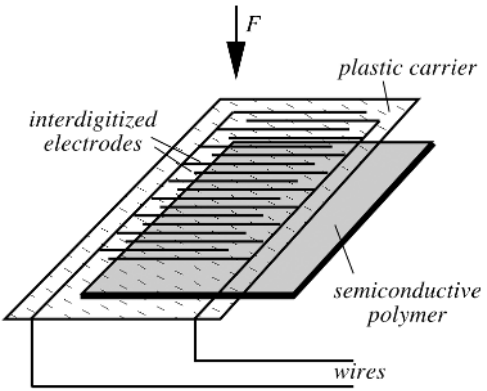


Fig. 9.9. Tactile sensor with a polymer FSR.

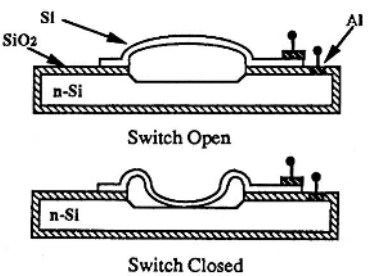


Fig. 9.10. Micromachined silicon threshold switch with trapped gas. (From Ref. [10].)

sealed cavity formed by wafer bonding is used to plastically deform a thin silicon membrane bonded over the cavity, creating a spherically shaped cap. The structure shown in Fig. 9.10 is fabricated by the micromachining technology of a silicon wafer. At normal room temperature and above a critical force, the upper electrode will buckle downward, making contact with the lower electrode.

Experiments have shown that the switch has a hysteresis of about 2 psi of pressure with a closing action near 13 psi. The closing resistance of the switch is on the order of 10 kΩ, which is usually low enough for the micropower circuits.

In another design, a vacuum, instead of pressurized gas, is used in a microcavity. This sensor, shown in Fig. 9.11 [11], has a silicon vacuum configuration, with a cold field-emission cathode and a movable diaphragm anode. The cathode is a sharp silicon tip. When a positive potential difference is applied between the tip and the anode, an electric field is generated, which allows electrons to tunnel from inside the cathode to the vacuum, if the field exceeds 5×10^7V/cm [12]. The field strength at the tip and the quantity of electrons emitted (emission current) are controlled by the anode potential. When an external force is applied, the anode deflects downward, thus changing the field and the emission current.

The emission current can be expressed through the anode voltage V as

$$I = V^2 a \exp\left(-\frac{b}{\beta V}\right), \tag{9.7}$$

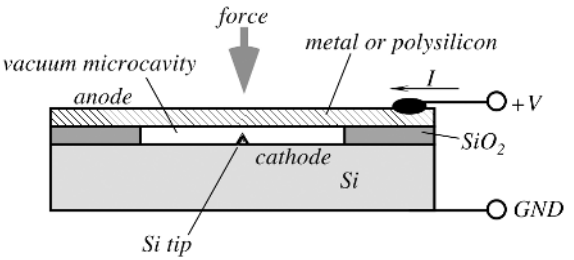


Fig. 9.11. Schematic of a vacuum diode force sensor. (Adapted from Ref. [11].)

where a and b are constants and β is the tip geometry factor, which depends on the distance between the anode and the cathode. To achieve a better sensitivity, the tip is fabricated with a radius of curvature of about 0.02 μm.

9.3 Piezoelectric Force Sensors

Although the tactile sensors that use piezoelectric effect as it was described earlier are not intended for the precision measurement of force, the same effect can be used quite efficiently for precision measurements. Piezoelectric effects can be used in both passive and active force sensors. It should be remembered, however, that a piezoelectric effect is, so to speak, an ac effect. In other words, it can convert a changing force into a changing electrical signal, whereas a steady-state force produces no electrical response. Yet, force can change some properties of a material that would affect an ac piezoelectric response when a sensor is supplied by an active excitation signal. One example of an active approach is shown in Fig. 9.4. However, for quantitative measurements, the applied force must be related to the mechanical resonant of the piezoelectric crystal. The basic idea behind the sensor's operation is that certain cuts of quartz crystal, when used as resonators in electronic oscillators, shift the resonant frequency upon being mechanically loaded. The equation describing the natural mechanical frequency spectrum of a piezoelectric oscillator is given by [13]

$$f_n = \frac{n}{2l}\sqrt{\frac{c}{\rho}}, \tag{9.8}$$

where n is the harmonic number, l is the resonance-determining dimension (e.g., the thickness of a relatively large thin plate or the length of a thin long rod), c is the effective elastic stiffness constant (e.g., the shear stiffness constant in the thickness direction of a plate or Young's modulus in the case of a thin rod), and ρ is the density of the crystal material.

The frequency shift induced by external force is due to nonlinear effects in the crystal. In the above equation, the stiffness constant c changes slightly with the applied stress. The effect of the stress on the dimension (strain) or the density is negligible. The minimal sensitivity to external force can occur when the squeezed dimension is

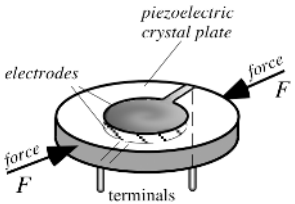


Fig. 9.12. A piezoelectric disk resonator as a diametric force sensor.

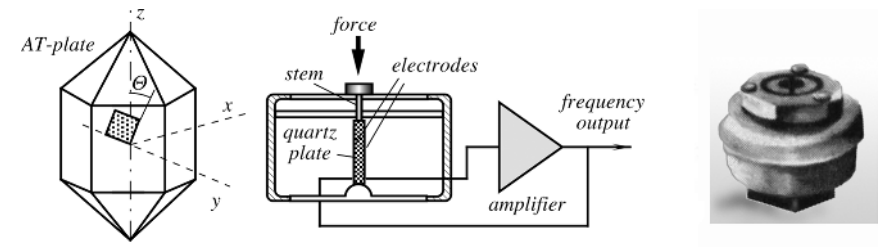


Fig. 9.13. Quartz force sensor: (A) AT-cut of a quartz crystal; (B) structure of the sensor; (C) the outside appearance. (Courtesy of Quartzcell, Santa Barbara, CA.)

aligned in certain directions for a given cut. These directions are usually chosen when crystal oscillators are designed, because their mechanical stability is important. However, in the sensor applications, the goal is just the opposite. For example, the diametric force has been used for a high-performance pressure transducer [14] (Fig. 9.12).

Another design of a sensor which operates over a relatively narrow range from 0 to 1.5 kg but with a good linearity and over 11-bit resolution is shown in Fig. 9.12. To fabricate the sensor, a rectangular plate is cut from the crystal such that only one edge is parallel to the x axis, and the face of the plate is cut at the angle of approximately $\Theta = 35°$ with respect to the z axis. This cut is commonly known as the AT-cut plate (Fig. 9.13A).

The plate is given surface electrodes for utilizing a piezoelectric effect (see Fig. 3.22 of Chapter 3), which are connected in a positive feedback of an oscillator (Fig. 9.13B). A quartz crystal oscillates at a fundamental frequency f_0 (unloaded) which shifts at loading by [15]

$$\Delta f = F \frac{K f_0^2 n}{l}, \tag{9.9}$$

where F is the applied force, K is a constant, n is the number of the overtone mode, and l is the size of the crystal. To compensate for frequency variations due to temperature effects, a double crystal can be employed, where one half is used for a temperature compensation. Each resonator is connected into its own oscillating circuit and the resulting frequencies are subtracted, thus negating a temperature effect. A commercial force sensor is shown in Fig. 9.13C.

A fundamental problem in all force sensors which use crystal resonators is based on two counterbalancing demands. On one hand, the resonator must have the highest

possible quality factor, which means the sensor must be decoupled from the environment and possibly should operate in vacuum. On the other hand, application of force or pressure requires a relatively rigid structure and a substantial loading effect on the oscillation crystal, thus reducing its quality factor.

This difficulty may be partly solved by using a more complex sensor structure. For instance, in a photolithographically produced double- and triple-beam structures [13,16], the so-called "string" concept is employed. The idea is to match dimensions of the oscillating element to the acoustic quarter-wavelength ($1/4\lambda$). The total wave reflection occurs at the supporting points through which the external force is coupled and the loading effect on the quality factor is significantly reduced.

References

1. Raman, V. V. The second law of motion and Newton equations. *Physics Teacher*, 1972.
2. Doebelin, E. O. *Measurement Systems: Applications and Design. McGraw-Hill*, New York, 1966.
3. Pallás-Areny, R. and Webster, J. G. *Sensors and Signal Conditioning*, 2nd ed. John Wiley & Sons, New York, 2001.
4. Holman J. P. *Experimental Methods for Engineers*. McGraw-Hill, New York, 1978.
5. Howe, R. T. Surface micromachining for microsensors and microactuators. *J. Vac. Sci. Technol. B*. 6(6), 1809–1813, 1988.
6. *Piezo Film Sensors Technical Manual*. Measurement Specialties, Inc., Norristown, PA, 1999; available at www.msiusa.com.
7. Fraden, J. Cardio-Respiration Transducer, U.S. Patent 4509527, 1985.
8. Del Prete, Z., Monteleone, L., and Steindler, R. A novel pressure array sensor based on contact resistance vartiation: metrological properties. *Rev. Sci. Instrum.* 72(3), 1548–1558, 2001.
9. Yates, B. A keyboard controlled joystick using force sensing resistor. *Sensors Magazine*, 39–39, 1991.
10. Huff, M.A., Nikolich, A.D., and Schmidt, M.A. A threshold pressure switch utilizing plastic deformation of silicon. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*, IEEE, New York, 1991, pp. 177–180.
11. Jiang, J.C., White, R.C., and Allen, P.K. Microcavity vacuum tube pressure sensor for robotic tactile sensing. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*. IEEE, New York, 1991, pp. 239–240
12. Brodie, I. Physical considerations in vacuum microelectronics devices. *IEEE Trans. Electron. Dev.*, 36, 2641, 1989.
13. Benes, E., Gröschl M., Burger W., and Schmid M. Sensors based on piezoelectric resonators. *Sensors Actuators A* 48, 1–21, 1995.

14. Karrer, E. and Leach J. A low range quartz pressure transducer. *ISA Trans*. 16, 90–98, 1977.
15. Corbett, J.P. Quatz steady-state force and pressure sensor. In: *Sensors Expo West Proceedings*. Helmers Publishing, Peterborough, NH, 1991.
16. Kirman, R.G. and Langdon, R.M. Force sensors. U.S. Patent 4,594,898. 1986.

This page intentionally left blank

10

Pressure Sensors

"To learn something new,
first, you must know something old."

—My physics teacher

10.1 Concepts of Pressure

The pressure concept was primarily based on the pioneering work of Evangelista Torricelli, who, for a short time, was a student of Galileo [1]. During his experiments with mercury-filled dishes, in 1643, he realized that the atmosphere exerts pressure on Earth. Another great experimenter, Blaise Pascal, in 1647, conducted an experiment, with the help of his brother-in-law Perier, on the top of the mountain Puy de Dome and at its base. He observed that pressure exerted on the column of mercury depends on elevation. He named the mercury-in-vacuum instrument they used in the experiment a *barometer*. In 1660, Robert Boyle stated his famous relationship: *The product of the measures of pressure and volume is constant for a given mass of air at fixed temperature*. In 1738, Daniel Bernoulli developed an impact theory of gas pressure to the point where Boyle's law could be deducted analytically. Bernoulli also anticipated the Charles–Gay–Lussac law by stating that pressure is increased by heating gas at a constant volume. For a detailed description of gas and fluid dynamics, the reader is referred to one of the many books on the fundamentals of physics. In this chapter, we briefly summarize the basics which are essential for the design and use of pressure sensors.

In general terms, matter can be classified into solids and fluids. The word *fluid* describes something which can flow. That includes liquids and gases. The distinction between liquids and gases are not quite definite. By varying pressure, it is possible to change liquid into gas and vice versa. It is impossible to apply pressure to fluid in any direction except normal to its surface. At any angle, except 90°, fluid will just slide over, or flow. Therefore, any force applied to fluid is tangential and the pressure exerted on boundaries is normal to the surface. For a fluid at rest, pressure

can be defined as the force F exerted perpendicularly on a unit area A of a boundary surface [2]:

$$p = \frac{dF}{dA}. \tag{10.1}$$

Pressure is basically a mechanical concept that can be fully described in terms of the primary dimensions of mass, length, and time. It is a familiar fact that pressure is strongly influenced by the position within the boundaries; however, at a given position, it is quite independent of direction. We note the expected variations in pressure with elevation:

$$dp = -w\, dh, \tag{10.2}$$

where w is the specific weight of the medium and h represents the vertical height.

Pressure is unaffected by the shape of the confining boundaries. Thus, a great variety of pressure sensors can be designed without concern for shape and dimensions. If pressure is applied to one of the sides of the surface confining a fluid or gas, the pressure is transferred to the entire surface without diminishing in value.

The kinetic theory of gases states that pressure can be viewed as a measure of the total kinetic energy of the molecules:

$$p = \frac{2}{3}\frac{\mathrm{KE}}{V} = \frac{1}{3}\rho C^2 = NRT, \tag{10.3}$$

where KE is the kinetic energy, V is the volume, C^2 is an average value of the square of the molecular velocities, ρ is the density, N is the number of molecules per unit volume, R is a specific gas constant, and T is the absolute temperature.

Equation (10.3) suggests that the pressure and density of compressible fluids (gases) are linearly related. The increase in pressure results in the proportional increase in density. For example, at 0°C and 1 atm pressure, air has a density of 1.3 kg/m^3, whereas at the same temperature and 50 atm of pressure, its density is 65 kg/m^3, which is 50 times higher. To the contrary, for liquids, the density varies very little over ranges of pressure and temperature. For instance, water at 0°C and 1 atm has a density of 1000 kg/m^3, whereas at 0°C and 50 atm, its density is 1002 kg/m^3, and at 100°C and 1 atm, its density is 958 kg/m^3.

Whether gas pressure is above or below the pressure of ambient air, we speak about overpressure or vacuum. Pressure is called relative when it is measured with respect to ambient pressure. It is called absolute when it is measured with respect to a vacuum at 0 pressure. The pressure of a medium may be static when it is referred to fluid at rest, or dynamic when it is referred to kinetic energy of a moving fluid.

10.2 Units of Pressure

The SI unit of pressure is the *pascal*: 1 Pa=1 N/m^2; that is, one pascal is equal to one newton of force uniformly distributed over 1 square meter of surface. Sometimes, in technical systems, *atmosphere* is used, which is denoted 1 atm. One atmosphere is the pressure exerted on 1 square centimeter by a column of water having a height of

1 meter at a temperature of +4°C and normal gravitational acceleration. A pascal can be converted into other units by the use of the following relationships (see also Table A.4 in the Appendix):

$$1\ \text{Pa} = 1.45 \times 10^{-4}\ \text{lb/in}^2 = 9.869 \times 10^{-6}\ \text{atm} = 7.5 \times 10^{-4}\ \text{cm Hg}.$$

For practical estimation, it is useful to remember that 0.1 mm H_2O is roughly equal to 1 Pa. In industry, another unit of pressure is often used. It is defined as pressure exerted by a 1-mm column of mercury at 0°C at normal atmospheric pressure and normal gravity. This unit is named after Torricelli and is called the torr. The ideal pressure of the Earth's atmosphere is 760 torr and is called the *physical atmosphere*:

$$1\ \text{atm} = 760\ \text{torr} = 101{,}325\ \text{Pa}.$$

The U.S. Customary System of units defines pressure as a pound per square inch (lb/sq in.) or psi. Conversion into SI systems is the following:

$$1\ \text{psi} = 6.89 \times 10^3 \text{Pa} = 0.0703\ \text{atm}.$$

A pressure sensor operating principle is based on the conversion of a result of the pressure exertion on a sensitive element into an electrical signal. Virtually in all cases, pressure results in the displacement or deformation of an element, having a defined surface area. Thus, a pressure measurement may be reduced to a measurement of a displacement or force, which results from a displacement. Thus, we recommend that the reader also becomes familiar with displacement sensors covered in Chapter 7 and force sensors of Chapter 9.

10.3 Mercury Pressure Sensor

A simple yet efficient sensor is based on the communicating vessels principle (Fig. 10.1). Its prime use is for the measurement of gas pressure. A U-shaped wire is immersed into mercury, which shorts its resistance in proportion with the height of mercury in each column. The resistors are connected into a Wheatstone bridge circuit, which remains in balance as long as the differential pressure in the tube is zero. Pressure is applied to one of the arms of the tube and disbalances the bridge, which results in the output signal. The higher the pressure in the left tube, the higher the resistance of the corresponding arm is and the lower the resistance of the opposite arm is. The output voltage is proportional to a difference in resistances ΔR of the wire arms which are not shunted by mercury:

$$V_{\text{out}} = V \frac{\Delta R}{R} = V \beta \Delta p. \tag{10.4}$$

The sensor can be directly calibrated in units of torr. Although simple, this sensor suffers from several drawbacks, such as necessity of precision leveling, susceptibility to shocks and vibration, large size, and contamination of gas by mercury vapors.[1]

[1] Note that this sensor can be used as an inclination sensor when pressures at both sides of the tube are equal.

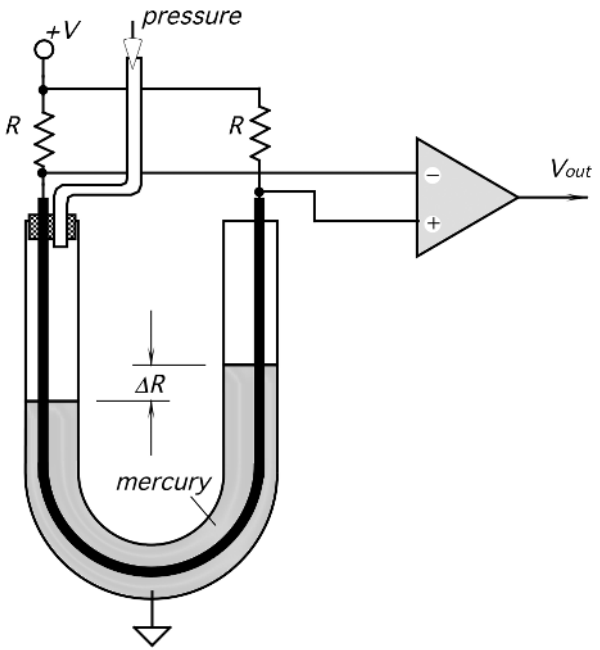


Fig. 10.1. Mercury-filled U-shaped sensor for measuring gas pressure.

10.4 Bellows, Membranes, and Thin Plates

In pressure sensors, a sensing element is a mechanical device which undergoes structural changes under strain. Historically, such devices were bourdon tubes (C-shaped, twisted, and helical), corrugated [3] and catenary diaphragms, capsules, bellows, barrel tubes, and other components whose shape changed under pressure.

A bellows (Fig. 10.2A) is intended for the conversion of pressure into a linear displacement which can be measured by an appropriate sensor. Thus, the bellows performs a first step in the conversion of pressure into an electrical signal. It is characterized by a relatively large surface area and, therefore, by a large displacement at low pressures. The stiffness of seamless metallic bellows is proportional to Young's modulus of the material and inversely proportional to the outside diameter and to the number of convolutions of the bellows. Stiffness also increases with roughly the third power of the wall thickness.

A popular example of pressure conversion into a linear deflection is a diaphragm in an aneroid barometer (Fig. 10.2B). A deflecting device always forms at least one wall of a pressure chamber and is coupled to a strain sensor (e.g., a strain gauge) which converts deflection into electrical signals. Currently, a great majority of pressure sensors are fabricated with silicon membranes by using micromachining technology.

A membrane is a thin diaphragm under radial tension S which is measured in Newtons per meter (Fig. 10.3B). The stiffness of bending forces can be neglected, as the thickness of the membrane is much smaller compared with its radius (at least 200 times smaller). When pressure is applied to one side of a membrane, it shapes it

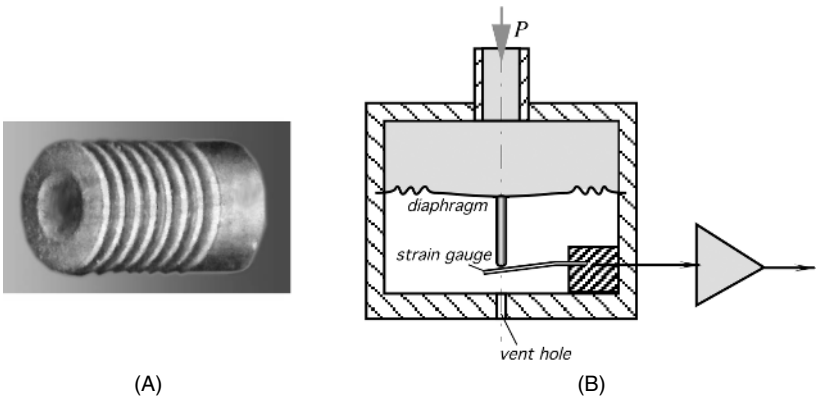


Fig. 10.2. (A) Steel bellows for a pressure transducer (fabricated by Servometer Corp., Cedar Grove, NJ); (B) metal corrugated diaphragm for conversion of pressure into linear deflection.

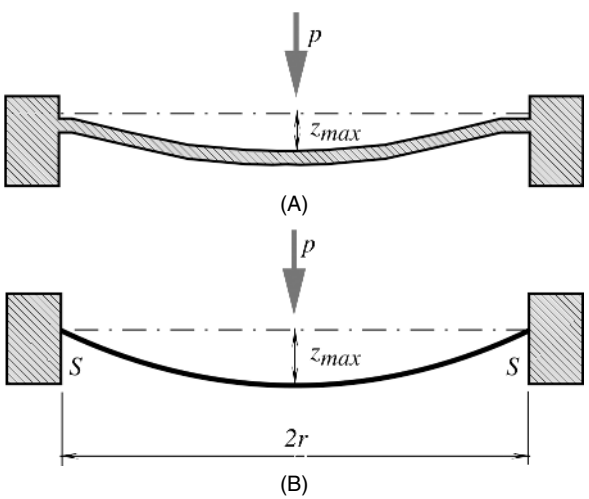


Fig. 10.3. Thin plate (A) and membrane (B) under pressure p.

spherically, like a soap bubble. At low-pressure p differences, the center deflection z_m and the stress σ_m are quasilinear functions of pressure[2]:

$$z_{\max} = \frac{r^2 p}{4S}, \tag{10.5}$$

$$\sigma_{\max} \approx \frac{S}{g}, \tag{10.6}$$

where r is the membrane radius and g is the thickness. Stress is generally uniform over the membrane area.

[2] Stress is measured in newtons per square meter.

For the membrane, the lowest natural frequency can be calculated from [4]

$$f_0 = \frac{1.2}{\pi r}\sqrt{\frac{S}{\rho g}}, \tag{10.7}$$

where ρ is the membrane material density. If the thickness of the membrane is not negligibly small (r/g ratio is 100 or less), the membrane is called a *thin plate* (Fig. 10.3A). If the plate is compressed between some kind of clamping rings, it exhibits a noticeable hysteresis due to friction between the thin plate and the clamping rings. A much better arrangement is a one-piece structure where the plate and the supporting components are fabricated of a single bulk of material.

For a plate, the maximum deflection is also linearly related to pressure:

$$z_{\max} = \frac{3(1-v^2)r^4 p}{16Eg^3}, \tag{10.8}$$

where E is Young's modulus (N/m^2) and v is Poisson's ratio. The maximum stress at the circumference is also a linear function of pressure:

$$\sigma_{\max} \approx \frac{3r^2 p}{4g^2}. \tag{10.9}$$

Equations (10.8) and (10.9) suggest that a pressure sensor can be designed by exploiting the membrane and thin plate deflections. The next question is: What physical effect should be used for the conversion of the deflection into an electrical signal? There are several options which we discuss in the following sections.

10.5 Piezoresistive Sensors

To make a pressure sensor, two essential components are required. They are the plate (membrane) having known area A and a detector which responds to applied force F [Eq. (10.1)]. Both of these components can be fabricated of silicon. A silicon-diaphragm pressure sensor consists of a thin silicon diaphragm as an elastic material [5] and a piezoresistive gauge resistors made by diffusive impurities into the diaphragm. Because of single-crystal silicon's superior elastic characteristics, virtually no creep and no hysteresis occur, even under strong static pressure. The gauge factor of silicon is many times stronger than that of thin metal conductors [6]. It is customary to fabricate strain gauge resistors connected as the Wheatstone bridge. The full-scale output of such a circuit is on the order of several hundred millivolts; thus, a signal conditioner is required for bringing the output to an acceptable format. Further, silicon resistors exhibit quite strong temperature sensitivity; therefore, a conditioning circuit should include temperature compensation.

When stress is applied to a semiconductor resistor, having initial resistance R, piezoresistive effect results in change in the resistance ΔR [7]:

$$\frac{\Delta R}{R} = \pi_1\sigma_1 + \pi_t\sigma_t, \tag{10.10}$$

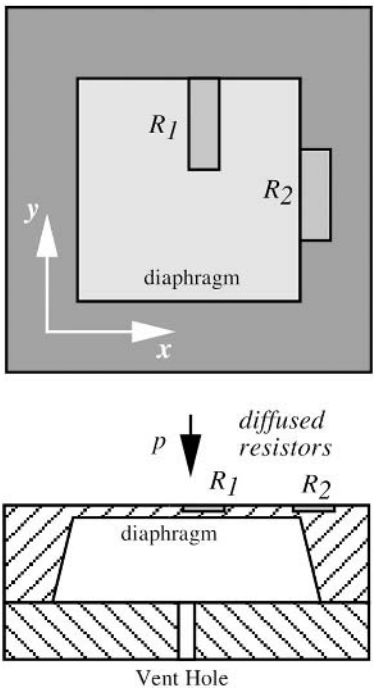


Fig. 10.4. Position of piezoresistors on a silicon diaphragm.

where π_1 and π_t are the piezoresistive coefficients in the longitudinal and transverse direction, respectively. Stresses in longitudinal and transverse directions are designated σ_1 and σ_t. The π coefficients depend on the orientation of resistors on the silicon crystal. Thus, for p-type diffused resistor arranged in the $\langle 110 \rangle$ direction or an n-type silicon square diaphragm with (100) surface orientation as shown in Fig. 10.4, the coefficients are approximately denoted as [7]

$$\pi_1 = -\pi_t = \frac{1}{2}\pi_{44}. \tag{10.11}$$

A change in resistivity is proportional to applied stress and, subsequently, to applied pressure. The resistors are positioned on the diaphragm in such a manner as to have the longitudinal and transverse coefficients of the opposite polarities; therefore, resistors change in the opposite directions:

$$\frac{\Delta R_1}{R_1} = -\frac{\Delta R_2}{R_2} = \frac{1}{2}\pi_{44}(\sigma_{1y} - \sigma_{1x}). \tag{10.12}$$

When connecting R_1 and R_2 in a half-bridge circuit and exciting the bridge with E, the output voltage V_{out} is

$$V_{\text{out}} = \frac{1}{4}E\pi_{44}(\sigma_{1y} - \sigma_{1x}). \tag{10.13}$$

As a result, pressure sensitivity a_p and temperature sensitivity of the circuit b_T can be found by taking partial derivatives:

$$a_p = \frac{1}{E}\frac{\partial V_{\text{out}}}{\partial p} = \frac{\pi_{44}}{4}\frac{\partial(\sigma_{1y} - \sigma_{1x})}{\partial p}, \tag{10.14}$$

$$b_T = \frac{1}{a_p}\frac{\partial a_p}{\partial T} = \frac{1}{\pi_{44}}\frac{\partial \pi_{44}}{\partial T}. \tag{10.15}$$

Since $\partial\pi_{44}/\partial T$ has a negative value, the temperature coefficient of sensitivity is negative; that is, sensitivity decreases at higher temperatures.

There are several methods of fabrication which can be used for silicon pressure sensor processing. In one method [8], the starting material is an n-type silicon substrate with (100) surface orientation. Piezoresistors with 3×10^{18}-cm^{-3} surface-impurity concentration are fabricated using a boron ion implantation. One of them (R_1) is parallel and the other is perpendicular to the ⟨110⟩ diaphragm orientation. Other peripheral components, like resistors and p-n junctions used for temperature compensation are also fabricated during the same implantation process as that for the piezoresistors. They are positioned in a thick-rim area surrounding the diaphragm. Thus, they are insensitive to pressure applied to the diaphragm.

Another approach of stress sensing was used in the Motorola MPX pressure sensor chip shown in Fig. 10.5 The piezoresistive element, which constitutes a strain gauge, is ion implanted on a thin silicon diaphragm. Excitation current is passed longitudinally through the resistor's taps 1 and 3, and the pressure that stresses the diaphragm is applied at a right angle to the current flow. The stress establishes a transverse electric field in the resistor that is sensed as voltage at taps 2 and 4. The single-element transverse voltage strain gauge can be viewed as the mechanical analog of a Hall effect device (Section 3.8 of Chapter 3). Using a single element eliminates the need to closely match the four stress- and temperature-sensitive resistors that form a Wheatstone

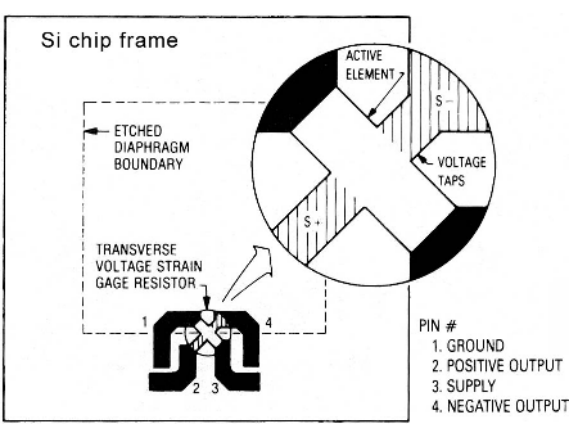


Fig. 10.5. Basic uncompensated piezoresistive element of Motorola MPX pressure sensor. (Copyright Motorola, Inc. Used with permission.)

bridge design. At the same time, it greatly simplifies the additional circuitry necessary to accomplish calibration and temperature compensation. Nevertheless, the single-element strain gauge electrically is analogous to the bridge circuit. Its balance (offset) does not depend on matched resistors, as it would be in a conventional bridge, but on how well the transverse voltage taps are aligned.

A thin diaphragm with 1-mm^2 area size may be formed by using one of the commonly used silicon etching solutions [e.g., hydrazine–water ($N_2H_4 \cdot H_2O$) anisotropic etchant]. A SiO_2 or Si_3N_4 layer serves as an etch mask and the protective layer on the bottom side of the wafer. The etching time is about 1.7 μm/min at 90°C in reflux solution. The final diaphragm thickness is achieved at about 30 μm.

Another method of diaphragm fabrication is based on the so-called silicon fusion bonding (SFB), where single crystal silicon wafers can be reliably bonded with near-perfect interfaces without the use of intermediate layers [9]. This technique allows the making of very small sensors which find use in catheter-tip transducers for medical in vivo measurements. The total chip area may be as much as eight times smaller than that of the conventional silicon-diaphragm pressure sensor. The sensor consists of two parts: the bottom and the top wafers (Fig. 10.6A). The bottom constraint wafer (substrate) is first anisotropically etched with a square hole which has the desirable dimensions of the diaphragm. The bottom wafer has a thickness about 0.5 mm and the diaphragm has side dimensions of 250 μm, so the anisotropic etch forms a pyramidal cavity with a depth of about 175 μm. The next step is SFB to a top wafer consisting of a *p*-type substrate with an *n*-type epi layer. The thickness of the epi layer corresponds to the desired final thickness of the diaphragm. Then, the bulk of the top wafer is removed by a controlled-etch process, leaving a bonded-on single crystal layer of silicon which forms the sensor's diaphragm. Next, resistors are ion implanted and contact vias are etched. In the final step, the constrain wafer is ground and polished back to the desired thickness of the device—about 140 μm. Despite the fact that the dimensions of the SFB chip are about half of those of the conventional chip, their pressure sensitivities are identical. A comparison of conventional and SFB technology is shown in Fig.

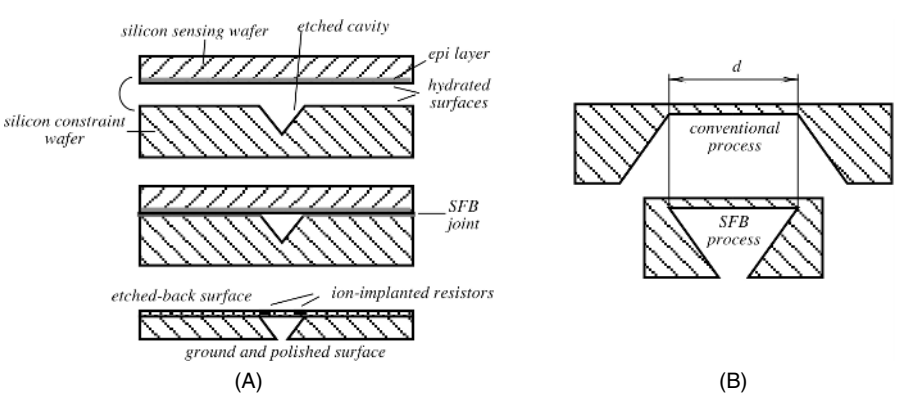


Fig. 10.6. Silicon fusion bonding method of a silicon membrane fabrications: (A) productions steps; (B) comparison of an SFB chip size with a conventionally fabricated diaphragm.

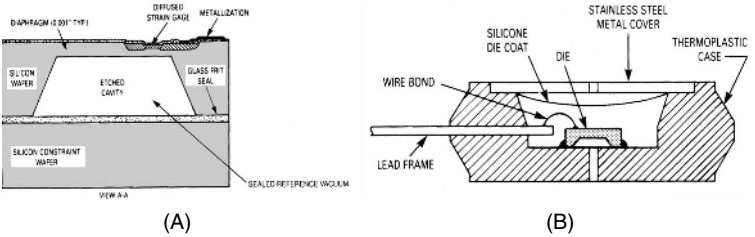


Fig. 10.7. Absolute (A) and differential (B) pressure sensor packagings. (Copyright Motorola, Inc. Used with permission.)

10.6B. For the same diaphragm dimensions and the same overall thickness of the chip, the SFB device is about 50% smaller.

Pressure sensors are usually available in three basic configurations that permit measurement of *absolute, differential*, and *gauge* pressures. Absolute pressure, such a barometric pressure, is measured with respect to a reference vacuum chamber. The chamber may be either external or it can be built directly into the sensor (Fig. 10.7A). A differential pressure, such as the pressure drop in a pressure-differential flowmeter, is measured by applying pressure to opposite sides of the diaphragm simultaneously. Gauge pressure is measured with respect to some kind of reference pressure. An example is a blood pressure measurement which is done with respect to atmospheric pressure. Thus, gauge pressure is a special case of a differential pressure. Diaphragm and strain gauge designs are the same for all three configurations; the packaging makes them different. For example, to make a differential or gauge sensor, a silicon die is positioned inside the chamber (Fig. 10.7B), which has two openings at both sides of the die. To protect them from a harsh environment, the interior of the housing is filled with a silicone gel which isolates the die surface and wire bonds while allowing the pressure signal to be coupled to the silicon diaphragm. A differential sensor may be incorporated into various porting holders (Fig. 10.8). Certain applications, such as a hot water hammer, corrosive fluids, and load cells, require physical isolation and hydraulic coupling to the chip-carrier package. It can be done with additional

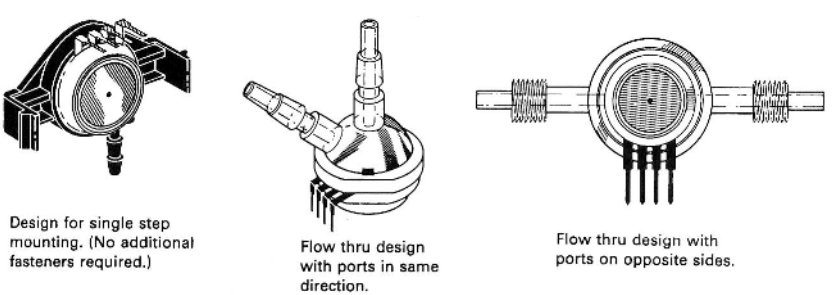

Design for single step mounting. (No additional fasteners required.)
Flow thru design with ports in same direction.
Flow thru design with ports on opposite sides.

Fig. 10.8. Examples of differential pressure packagings. (Copyright Motorola, Inc. Used with permission.)

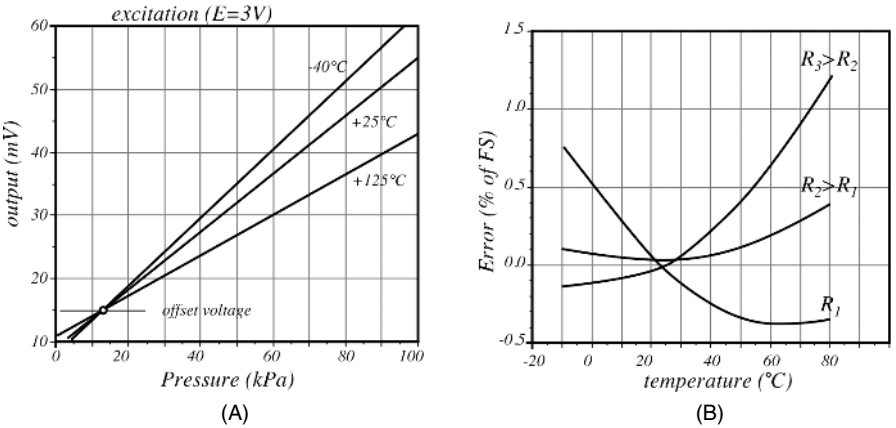


Fig. 10.9. Temperature characteristics of a piezoresistive pressure sensor: (A) transfer function at three different temperatures; (B) full-scale errors for three values of compensating resistors.

diaphragms and bellows. In either case, silicone oil, such as Dow Corning DS200, can be used to fill the air cavity so that system frequency response is not degraded.

All silicon-based sensors are characterized by temperature dependence. The temperature coefficient of sensitivity b_T as defined by Eq. (10.15) is usually negative, and for the accurate pressure sensing, it must be compensated for. Typical methods of temperature compensation of bridge circuits are covered in Section 5.7.3 of Chapter 5. Without the compensation, the sensor's output voltage may look like the one shown in Fig. 10.9A for three different temperatures.

In many applications, a simple yet efficient temperature compensation can be accomplished by adding to the sensor either a series or parallel temperature stable resistor. By selecting an appropriate value of the resistor, the sensor's output can be tailored to the desirable operating range (Fig. 10.9B). Whenever a better temperature correction over a broad range is required, more complex compensation circuits with temperature detectors can be employed. A viable alternative is a software compensation where the temperature of the pressure transducer is measured by an imbedded temperature sensor. Both data from the pressure and temperature sensors are relayed to the processing circuit where numerical compensation is digitally performed.

10.6 Capacitive Sensors

A silicon diaphragm can be used with another pressure-to-electric output conversion process: in a capacitive sensor. Here, the diaphragm displacement modulates capacitance with respect to the reference plate (backplate). This conversion is especially effective for the low-pressure sensors. An entire sensor can be fabricated from a solid piece of silicon, thus maximizing its operational stability. The diaphragm can be designed to produce up to 25% capacitance change over the full range which

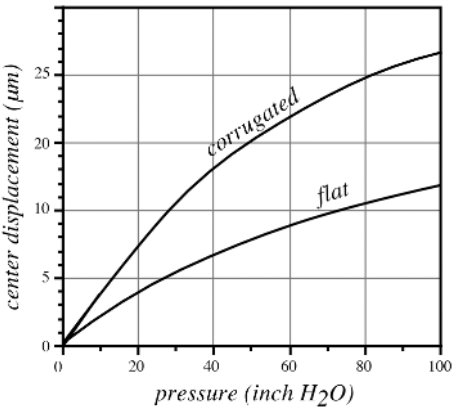


Fig. 10.10. Central deflection of flat and corrugated diaphragms of the same sizes under the in-plate tensile stresses.

makes these sensors candidates for direct digitization (see Section 5.5 of Chapter 5). Whereas a piezoresistive diaphragm should be designed to maximize stress at its edges, the capacitive diaphragm utilizes a displacement of its central portion. These diaphragms can be protected against overpressure by including mechanical stops close to either side of the diaphragm (for a differential pressure sensor). Unfortunately, in the piezoresistive diaphragms, the same protection is not quite effective because of small operational displacements. As a result, the piezoresistive sensors typically have burst pressures of about 10 times the full-scale rating, whereas capacitive sensors with overpressure stops can handle 1000 times the rated full-scale pressure. This is especially important for the low-pressure applications, where relatively high-pressure pulses can occur.

While designing a capacitive pressure sensor, for good linearity it is important to maintain flatness of the diaphragm. Traditionally, these sensors are linear only over the displacements which are much less than their thickness. One way to improve the linear range is to make a diaphragm with groves and corrugations by applying micromachining technology. Planar diaphragms are generally considered more sensitive than the corrugated diaphragms with the same size and thickness. However, in the presence of the in-plane tensile stresses, the corrugations serve to release some of the stresses, thus resulting in better sensitivity and linearity (Fig. 10.10).

10.7 VRP Sensors

When measuring small pressures, the deflection of a thin plate or a diaphragm can be very small. In fact, it can be so small that the use of strain gauges attached to or imbedded into the diaphragm becomes impractical due to the low output signal. One possible answer to the problem may be a capacitive sensor for which a diaphragm deflection is measured by its relative position to a reference base rather than by the internal strain in the material. Such sensors were described earlier. Another solution which is especially useful for very low pressures is a magnetic sensor. A *variable*

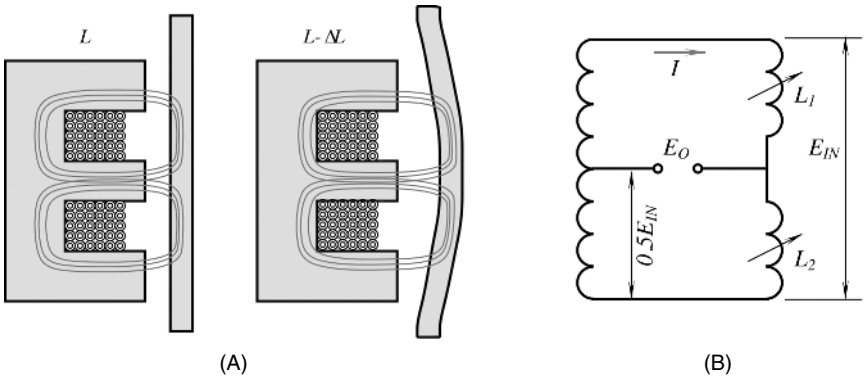


Fig. 10.11. Variable reluctance pressure sensor: (A) basic principle of operation; (B) an equivalent circuit.

reluctance pressure (VRP) sensor uses a magnetically conductive diaphragm to modulate the magnetic resistance of a differential transformer. The operation of the sensor is very close to that of the magnetic proximity detectors described in Chapter 5. Figure 10.11A illustrates a basic idea behind the magnetic flux modulation. The assembly of an E-shaped core and a coil produces a magnetic flux whose field lines travel through the core, the air gap and the diaphragm. The permeability of the E-core magnetic material is at least 1000 times higher than that of the air gap [10], and, subsequently, its magnetic resistance is lower than the resistance of air. Since the magnetic resistance of the air gap is much higher than the resistance of the core, it is the gap which determines the inductance of the core-coil assembly. When the diaphragm deflects, the air gap increases or decreases depending on the direction of a deflection, thus causing the modulation of the inductance.

To fabricate a pressure sensor, a magnetically permeable diaphragm is sandwiched between two halves of the shell (Fig. 10.12). Each half incorporates an E-core/coil assembly. The coils are encapsulated in a haÈd compound to maintain maximum stability under even very high pressure. Thin pressure cavities are formed at both sides of the diaphragm. The thickness of the diaphragm defines a full-scale operating range; however, under most circumstances, total deflection does not exceed 25–30 μm, which makes this device very sensitive to low pressures. Further, due to thin pressure cavities, the membrane is physically prevented from excessive deflection under the overpressure conditions. This makes VRP sensors inherently safe devices. When excited by an ac current, a magnetic flux is produced in each core and the air gaps by the diaphragm. Thus, the sensors contain two inductances and can, therefore, be thought of as half of a variable reluctance bridge where each inductance forms one arm of the bridge (Fig. 10.11B). As a differential pressure across the diaphragm is applied, the diaphragm deflects, one side decreasing and the other increasing, and the air-gap reluctances in the electromagnetic circuit change proportionally to the differential pressure applied. A full-scale pressure on the diaphragm, although very small, will produce a large output signal that is easily differentiated from noise.

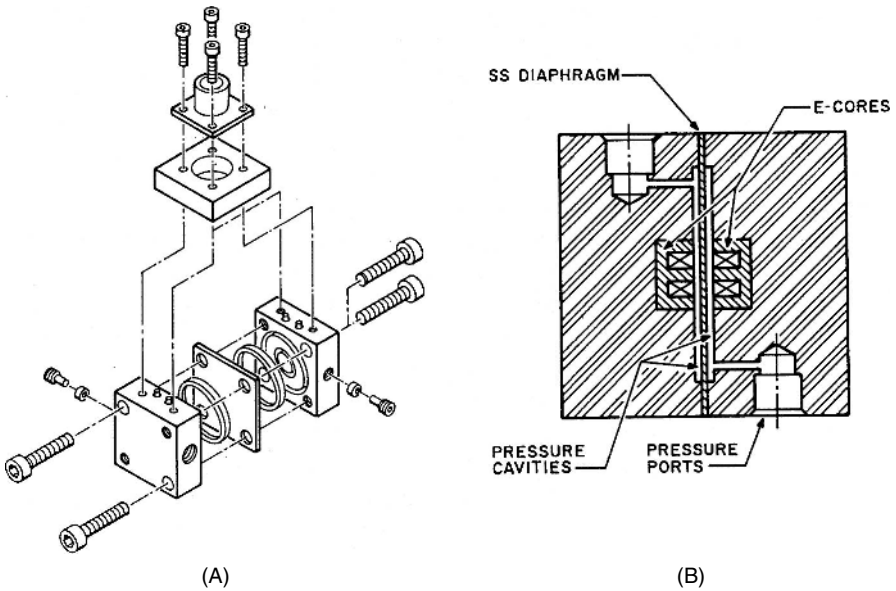


Fig. 10.12. Construction of a VRP sensor for low-pressure measurements: (A) assembly of the sensor; (B) double E-core at both sides of the cavity.

The VRP sensor's output is proportional to the reluctance in each arm of the inductive Wheatstone bridge that uses the equivalent inductive reactances $x_{1,2}$ as the active elements. The inductance of a coil is determined by the number of turns and the geometry of the coil. When a magnetically permeable material is introduced into the field flux, it forms a low-resistance path-attracting magnetic field. This alters the coil's self-inductance. The inductance of the circuit, and subsequently its reactance, is inversely proportional to the magnetic reluctance; that is, $x_{1,2} = k/d$ where k is a constant and d is the gap size. When the bridge is excited by a carrier, the output signal across the bridge becomes amplitude modulated by the applied pressure. The amplitude is proportional to the bridge imbalance, and the phase of the output signal changes with the direction of the imbalance. The ac signal can be demodulated to produce a dc response.

10.8 Optoelectronic Sensors

When measuring low-level pressures or, to the contrary, when thick membranes are required to enable a broad dynamic range, a diaphragm displacement may be too small to assure sufficient resolution and accuracy. In addition, most of piezoresistive sensors, and some capacitive, are quite temperature sensitive, which requires an additional thermal compensation. An optical readout has several advantages over other technologies, namely a simple encapsulation, small temperature effects, high

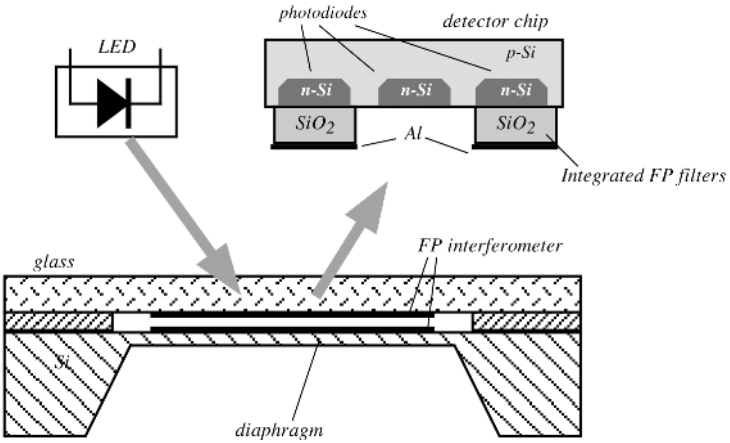


Fig. 10.13. Schematic of an optoelectronic pressure sensor operating on the interference phenomenon. (Adapted from Ref. [12].)

resolution, and high accuracy. Especially promising are the optoelectronic sensors operating with the light interference phenomena [11]. Such sensors use a Fabry–Perot (FP) principle of measuring small displacements (covered in more detail in Section 7.5 of Chapter 7). A simplified circuit of one such a sensor is shown in Fig. 10.13.

The sensor consists of the following essential components: a passive optical pressure chip with a membrane etched in silicon, a light-emitting diode (LED), and a detector chip [12]. A pressure chip is similar to a capacitive pressure sensor as described earlier, except that a capacitor is replaced by an optical cavity forming a Fabry–Perot interferometer [13] measuring the deflection of the diaphragm. A back-etched single-crystal diaphragm on a silicon chip is covered with a thin metallic layer, and a glass plate is covered with a metallic layer on its backside. The glass is separated from the silicon chip by two spacers at a distance w. Two metallic layers form a variable-gap FP interferometer with a pressure-sensitive movable mirror (on the membrane) and a plane-parallel, stationary, fixed half-transparent mirror (on the glass). A detector chip contains three p-n-junction photodiodes. Two of them are covered with integrated optical FP filters of slightly different thicknesses. The filters are formed as first surface silicon mirrors, coated with a layer of SiO_2 and thin metal (Al) mirrors on their surfaces. An operating principle of the sensor is based on the measurement of a wavelength modulation of the reflected and transmitted light depending on the width of the FP cavity. The reflection and transmission from the cavity is almost a periodic function in the inverse wavelength, $1/\lambda$, of the light with a period equal to $1/2w$. Because w is a linear function of the applied pressure, the reflected light is wavelength modulated.

The detector chip works as a demodulator and generates electrical signals representing the applied pressure. It performs an optical comparison of the sensing cavity of the pressure sensor with a virtual cavity formed by the height difference between two FP filters. If both cavities are the same, the detector generates the maximum pho-

tocurrent, and when the pressure changes, the photocurrent is cosine modulated with a period defined by half the mean wavelength of the light source. The photodiode without the FP filter serves as a reference diode, which monitors the total light intensity arriving at the detector. Its output signal is used for the ratiometric processing of the information. Because the output of the sensor is inherently nonlinear, a linearization by a microprocessor is generally required. Similar optical pressure sensors can be designed with fiber optics, which makes them especially useful for remote sensing where radio-frequency interferences present a serious problem [14].

10.9 Vacuum Sensors

Measurement of very low pressures is important for the processing of the microelectronic wafers, optical components, chemistry and other industrial applications. It also vital for the scientific studies, for instance, in space exploration. In general, *vacuum* means pressure below atmospheric, but usually the term is used with respect to a near absence of gas pressure. True vacuum is never attained. Even the intrastellar space is not entirely free of matter.

Vacuum can be measured as negative pressure compared to the atmospheric pressure by conventional pressure sensors, yet this is not quite efficient. Conventional pressure sensors do not resolve extremely low concentrations of gas due to the poor signal-to-noise ratio. Whereas the pressure sensors in most cases employ some kind of membrane and a displacement (deflection) transducer, special vacuum sensors operate on different principles. They rely on some physical properties of gaseous molecules that are related to the number of such molecules per volume of space. These properties may be a thermal conductivity, viscosity, ionization, and others. Here, we briefly describe some popular sensor designs.

10.9.1 Pirani Gauge

The Pirani vacuum gauge is a sensor that measures pressure through the thermal conductivity of gas. It is one of the oldest vacuum sensors. The simplest version of the gauge contains a heated plate. The measurement is done by detecting of amount of heat lost from the plate that depends on gas pressure. Operation of the Pirani gauge is based on the pioneering works by Marian Von Smoluchowski [15]. He established that when an object is heated, thermal conductivity to the surrounding objects is governed by

$$G = G_0 + G_g = G_s + G_r + ak\frac{P\,P_T}{P + P_T}, \tag{10.16}$$

where G_s is thermal conductivity via the solid supporting elements, G_e is the radiative heat transfer, a is the area of a heated plate, k is a coefficient related to gas properties and P_T is a transitional pressure which is the maximum pressure that can be measured. Figure 10.14A illustrates different factors that contributes to a thermal loss from a heated plate. If the solid conductive and radiative loss is accounted for, the gas

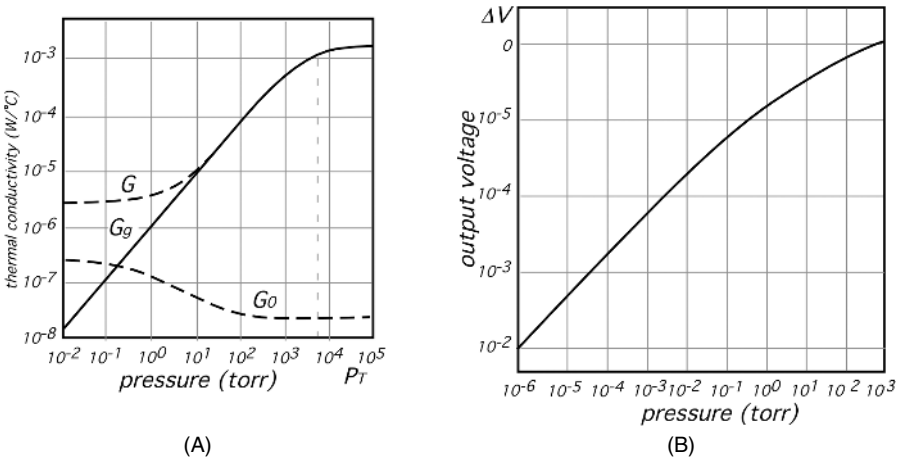


Fig. 10.14. (A) Thermal conductivities from a heated plate; (B) transfer function of a Pirani vacuum gauge.

conductivity G_g goes linearly down to absolute vacuum. The trick is to minimize the interfering factors that contribute to G_0. This can be achieved by use of both the heated plate that is suspended with a minimal thermal contact with the sensor housing and by the differential technique that to a large degree cancels the influence of G_0.

There are several designs of the Pirani gauge that are used in vacuum technologies. Some use two plates with different temperatures and the amount of power spent for heating is the measure of gas pressure. The other use a single plate that measures thermal conductivity of gas by heat loss to the surrounding walls. Temperature measurement is usually done with either a thermocouple or platinum resistive temperature detector (RTD). Figure 10.15 illustrates one version of the gauge that employs a thermal balance (differential) technique. The sensor chamber is divided into two identical sections where one is filled with gas at a reference pressure (say 1 atm $= 760$ torr) and the other is connected to the vacuum that is to be measured. Each chamber contains a heated plate that is supported by the tiny links to minimize a conductive heat transfer through solids. Both chambers are preferably of the same shape, size, and construction so that the conductive and radiative heat loss would be nearly identical. The better the symmetry, the better is the cancellation of the spurious thermal conductivity G_0. The heaters on the plates are warmed up by electric current. In this particular design, each heater is a thermistor with a negative temperature coefficient (NTC) (see Chapter 16). Resistances of the thermistor are equal and relatively low to allow for a Joule self-heating (Fig. 16.11 of Chapter 16). The reference thermistor S_r is connected into a self-balancing bridge that also includes resistors R_r, R_1, and R_2 and an operational amplifier. The bridge automatically sets the temperature of S_r on a constant level T_r that is defined by the bridge resistors and is independent of ambient temperature. Note that the bridge is balanced by both the negative and positive feedbacks to the bridge arms. Capacitor C keeps the circuit from oscillating,

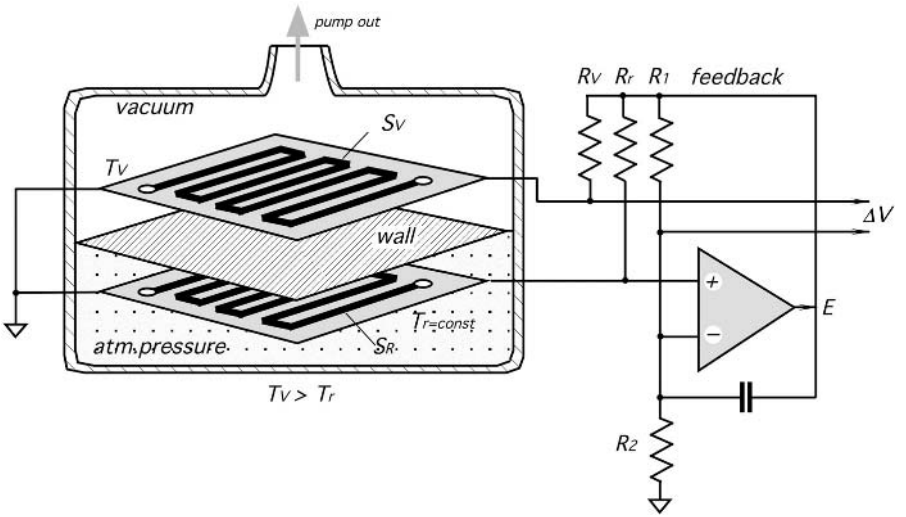


Fig. 10.15. Pirani vacuum gauge with NTC thermistors operating in self-heating mode.

The same voltage E that feeds the reference plate is applied to the thermistor S_v on the sensing plate via $R_v = R_r$. The output voltage ΔV is taken differentially from the sensing thermistor and the bridge. The shape of the transfer function is shown in Fig. 10.14B. A vacuum sensor often operates with gases that may contaminate the sensing plates so the appropriate filters must be employed.

10.9.2 Ionization Gauges

This sensor resembles a vacuum tube that was used as an amplifier in the old-fashioned radio equipment. The ion current between the plate and the filament (Fig. 10.16A) is a nearly linear function of molecular density (pressure) [16,17]. The vacuum gauge tube has a reversed connection of voltages: The positive high voltage is applied to a grid and negative lower voltage is connected to the plate. The output is the ion current i_p collected by the plate that is proportional to pressure and the electron current i_g of the grid. Presently, a further improvement of this gauge is the so-called Bayard–Alpert vacuum sensor [18]. It is more sensitive and stable at a much lower pressure. Its operating principle is the same as a vacuum tube gauge, except that the geometry is different—the plate is substituted by the a wire surrounded by a grid and the cathode filament is outside (Fig. 10.16B).

10.9.3 Gas Drag Gauge

The gas molecules interact with a moving body. This is the basic idea behind the spinning-rotor gauge [19]. In the current implementation of the sensor, a small steel ball having a diameter of 4.5 mm is magnetically levitated (Fig. 10.16C) inside a vacuum chamber and spinning with a rate of 400 Hz. The ball magnetic moment

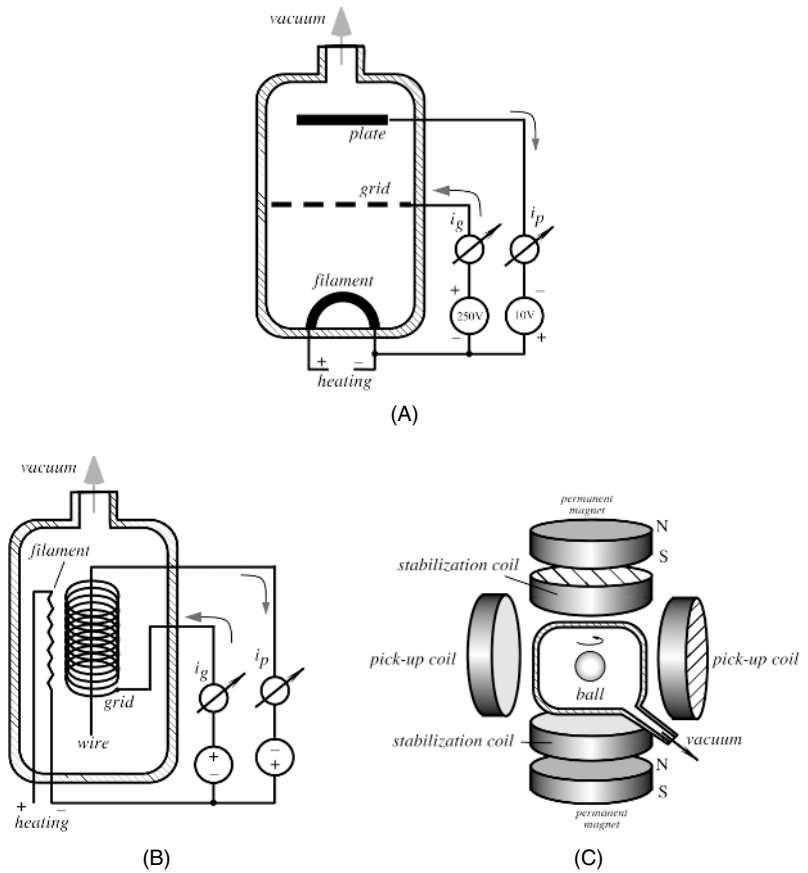


Fig. 10.16. Ionization vacuum gauge (A), Bayard–Alpert gauge (B), and gas drag gauge (C).

induces a signal in a pickup coil. The gas molecules exert drag on the ball and slow its rate of rotation:

$$P = \frac{\pi \rho a \bar{c}}{10\sigma_{\text{eff}}} \left(\frac{-\omega' - RD - 2\alpha T'}{\omega} \right) \tag{10.17}$$

where ρ and a are the density and radius of the ball, respectively, ω'/ω is the fractional rate of slowing of rotation, $\bar{c}$ is the mean gas molecular velocity, α is the coefficient of expansion of the ball, and T' is the ball's temperature [20].

References

1. Benedict, R. P. *Fundamentals of Temperature, Pressure, and Flow Measurements*, 3rd ed. John Wiley & Sons, New York, 1984.
2. Plandts, L. *Essentials of Fluid Dynamics*. Hafner, New York, 1952.

3. Di Giovanni, M. *Flat and Corrugated Diaphragm Design Handbook*. Marcel Dekker, New York, 1982.
4. Neubert, H. K. P. *Instrument Transducers. An Introduction to Their Performance and Design*, 2nd ed., Clarendon Press, Oxford, 1975.
5. Clark, S. K. and Wise, K. D. Pressure sensitivity in anisotropically etched thin-diaphragm pressure sensor. *IEEE Trans. Electron Dev.*, ED-26, 1887–1896, 1979.
6. Tufte, O. N., Chapman, P.W. and Long, D. Silicon diffused-element piezoresistive diaphragm. *J. Appl. Phys.* 33, 3322–3327, 1962 .
7. Kurtz, A. D. and Gravel, C. L. Semiconductor transducers using transverse and shear piezoresistance. Proc. 22nd ISA Conference, 1967.
8. Tanigawa, H., Ishihara, T., Hirata, M., and Suzuki K. MOS integrated silicon pressure sensor. *IEEE Trans. Electron Dev.* ED-32(7), 1191–1195, 1985.
9. Petersen, K., Barth, P., Poydock, J., Brown, J., Mallon, J., Jr., and Bryzek, J. Silicon fusion bonding for pressure sensors. Record of the IEEE Solid-State Sensor and Actuator Workshop, 1988, pp. 144–147.
10. Proud, R. VRP transducers for low-pressure measurement. *Sensors Magazine*, 20–22, 1991.
11. Wolthuis, R., A., Mitchell, G.L., Saaski, E., Hratl, J.C., and Afromowitz, M.A. Development of medical pressure and temperature sensors employing optical spectral modulation. *IEEE Trans. Biomed. Eng.*, 38(10), 974–981, 1991.
12. Hälg, B. A silicon pressure sensor with an interferometric optical readout. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*. IEEE, New York, 1991, pp. 682–684.
13. Vaughan, J.M. *The Fabry–Perot Interferometers*. Adam Hilger, Bristol, 1989.
14. Saaski, E.W., Hartl, J.C., and Mitchell, G.L. A fiber optic sensing system based on spectral modulation. Paper #86-2803, ISA, 1989.
15. Von Smoluchowski, M. *Ann-Phys.* 35, 983, 1911.
16. Buckley, O.E. *Proc. Natl. Acad. Sci.*, USA 2, 683, 1916.
17. Leck, J.H. *Pressure Measurement in Vacuum Systems*. Chapman & Hall., London, 1957, pp. 70–74.
18. Bayard, R.T. and Alpert, D. *Rev. Sci. Instrum.* 21, 571, 1950.
19. Fremery, J.K. *Vacuum* 32, 685, 1946.
20. Goehner, R., Drubetsky, E., Brady, H.M, and Bayles, W.H., Jr. Vacuum measurement. In: *Mechanical Variables Measurement*. Webster, ed. CRC Press, Boca Raton, FL, 2000.

11

Flow Sensors

It's a simple task to make a complex system,
It's a complex task to make a simple system

11.1 Basics of Flow Dynamics

One of the fundamentals of physics is that mass is a conserved quantity. It cannot be created or destroyed. In the absence of sources or sinks of mass, its quantity remains constant regardless of boundaries. However, if there is influx or outflow of mass through the boundaries, the sum of influx and efflux must be zero. Whatever mass comes in, it must go out. When both are measured over the same interval of time, mass entering the system (M_{in}) is equal to mass leaving the system (M_{out}) [1]. Therefore,

$$\frac{dM_{\text{in}}}{dt} = \frac{dM_{\text{out}}}{dt}. \tag{11.1}$$

In mechanical engineering, moving media whose flow is measured are liquids (water, oil, solvents, gasoline, etc.), air, gases (oxygen, nitrogen, CO, CO_2, methane CH_4, water vapor, etc.).

In a steady flow, the velocity at a given point is constant in time. We can draw a stream line through every point in a moving medium (Fig. 11.1A). In steady flow, the line distribution is time independent. A velocity vector is tangent to a stream line in every point z. Any boundaries of flow which envelop a bundle of stream lines is called a *tube of flow*. Because the boundary of such a tube consists of stream lines, no fluid (gas) can cross the boundary of a tube of flow and the tube behaves something like a pipe of some shape. The flowing medium can enter such a pipe at one end, having cross section A_1 and exit at the other through cross section A_2. The velocity of a moving material inside of a tube of flow will, in general, have different magnitudes at different points along the tube.

The volume of moving medium passing a given plane (Fig. 11.1B) in a specified time interval Δt is

$$\Lambda = \frac{V}{\Delta t} = \int \frac{\Delta x \, dA}{\Delta t} = \int v \, dA, \tag{11.2}$$

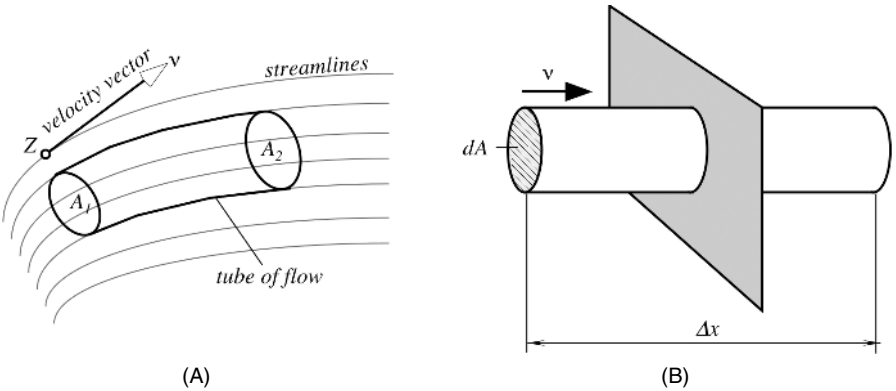


Fig. 11.1. Tube of flow (A) and flow of a medium through a plane (B).

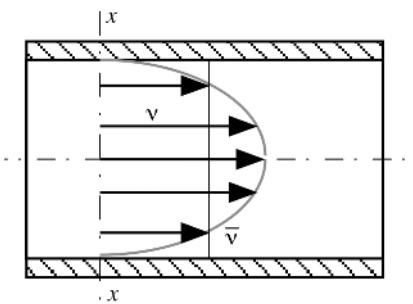


Fig. 11.2. Profile of the velocity of flow in a pipe.

where v is the velocity of moving medium, which must be integrated over area A, and Δx is the displacement of volume V. Figure 11.2 shows that the velocity of a liquid or gas in a pipe may vary over the cross section. It is often convenient to define an average velocity

$$v_a = \frac{\int v \, dA}{A}. \tag{11.3}$$

When measuring the velocity by a sensor whose dimensions are substantially smaller than the pipe size, one should be aware of the possibility of erroneous detection of either too low or too high velocity, whereas the average velocity, v_a, is somewhere in between. A product of the average velocity and a cross-sectional area is called the *flux* or *flow rate*. Its SI unit is cubic meters per second. The U.S. Customary System unit is cubic feet per second. The flux can be found by rearranging Eq. (11.3):

$$Av_a = \int v \, dA. \tag{11.4}$$

What a flow sensor usually measures is v_a. Thus, to determine the flow rate, the cross-section area of tube of flow A must be known, otherwise the measurement is meaningless.

The measurement of flow is rarely conducted for the determination of a displacement of volume. Usually, what is needed is to determine the flow of mass rather than volume. Of course, when dealing with virtually incompressible fluids (water, oil, etc.), either volume or mass can be used. A relationship between mass and volume for a incompressible material is through density ρ

$$M = \rho V. \tag{11.5}$$

The densities of some materials are given in Table A.12 (Appendix). The rate of mass flow is defined as

$$\frac{dM}{dt} = \rho A \overline{v} \tag{11.6}$$

The SI unit for mass flow is kilogram per second and the U.S. Customary System unit is pounds per second. For a compressible medium (gas), either mass flow or volume flow at a given pressure should be specified.

There is a great variety of sensors that can measure flow velocity by determining the rate of displacement of either mass or volume. Whichever sensor is used, inherent difficulties of the measurement make the process a complicated procedure. It is necessary to take into consideration many of the natural characteristics of the medium, its surroundings, barrel and pipe shapes and materials, medium temperature and pressure, and so forth. When selecting any particular sensor for the flow measurement, it is advisable to consult with the manufacturer's specifications and very carefully consider the application recommendations for a particular sensor. In this book, we do not cover such traditional flow measurement systems as turbine-type meters. It is of interest to us to consider sensors without moving components which introduce either no or little restriction into the flow.

11.2 Pressure Gradient Technique

A fundamental equation in fluid mechanics is *Bernoulli equation* which is strictly applicable only to steady flow of nonviscous, incompressible medium:

$$p + \rho \left(\frac{1}{2} v_a^2 + gy \right) = \text{const}, \tag{11.7}$$

where p is the pressure in a tube of flow, $g = 9.80665\ \text{m/s}^2 = 32.174\ \text{ft/s}^2$ is the gravity constant, and y is the height of the medium's displacement. Bernoulli's equation allows us to find fluid velocity by measuring pressures along the flow.

The pressure gradient technique (of flow measurement) essentially requires the introduction of a flow resistance. Measuring the pressure gradient across a known resistor allows one to calculate a flow rate. The concept is analogous to Ohm's law: Voltage (pressure) across a fixed resistor is proportional to current (flow). In practice, the restricting elements which cause flow resistances are orifices, porous plugs, and Venturi tubes (tapered profile pipes). Figure 11.3 shows two types of flow resistor. In the first case, it is a narrow in the channel; in the other case, there is a porous

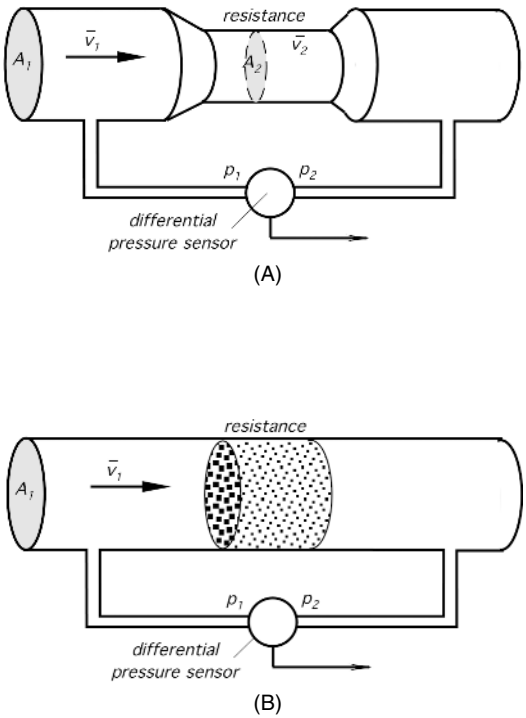


Fig. 11.3. Two types of flow resistor: a narrow channel (A) and a porous plug (B).

plug which somewhat restricts the medium flow. A differential pressure sensor is positioned across the resistor. When moving mass enters the higher-resistance area, its velocity increases in proportion to the resistance increase:

$$v_{1a} = v_{2a} R. \tag{11.8}$$

The Bernoulli equation defines differential pressure as[1]

$$\Delta p = p_1 - p_2 = \frac{\rho}{2}(v_{2a}^2 - v_{1a}^2) = k\frac{\rho}{2}v_{2a}^2(1 - R^2), \tag{11.9}$$

where k is the correction coefficient which is required because the actual pressure p_2 is slightly lower than the theoretically calculated pressure. From Eq. (11.9), the average velocity can be calculated as

$$v_{2a} = \frac{1}{\sqrt{k(1 - R^2)}}\sqrt{\frac{2}{\rho}\Delta p}. \tag{11.10}$$

[1] It is assumed that both pressure measurements are made at the same height ($y = 0$), which is usually the case.

To determine the mass flow rate per unit time, for a incompressible medium, Eq. (11.10) is simplified to

$$q = \xi A_2 \sqrt{\Delta p}, \tag{11.11}$$

where ξ is a coefficient which is determined through calibration. The calibration must be done with a specified liquid or gas over an entire operating temperature range; thus, the value of ξ may be different at different temperatures. It follows from the above that the pressure gradient technique essentially requires the use of either one differential pressure sensor or two absolute sensors. If a linear representation of the output signal is required, a square root extraction must be used. The root extraction can be performed in a microprocessor by using one of the conventional computation techniques. An advantage of the pressure gradient method is in the absence of moving components and use of standard pressure sensors which are readily available. A disadvantage is in the restriction of flow by resistive devices.

11.3 Thermal Transport Sensors

A good method for measuring flow would be by somehow marking the flowing medium and detecting the movement of the mark. For example, a mark can be a floating object that can move with the medium while being stationary with respect to the medium. The time which it would take the object to move with the flow from one position to another could be used for the calculation of the flow rate. Such an object may be a float, radioactive element, or dye which changes optical properties (e.g., color) of flowing medium. Also, the mark can be a different gas or liquid whose concentration and rate of dilution can be detectable by appropriate sensors.

In medicine, a dye dilution method of flow measurement is used for studies in hemodynamics. In most instances, however, placing any foreign material into the flowing medium is either impractical or forbidden for some other reasons. An alternative would be to change some physical properties of the moving medium and to detect the rate of displacement of a changed portion or rate of its dilution. Usually, the best physical property that can be easily modified without causing undesirable effects is temperature.

Figure 11.4A shows a sensor which is called a thermoanemometer. It is composed of three small tubes immersed into a moving medium. Two tubes contain temperature detectors R_0 and R_s. The detectors are thermally coupled to the medium and are thermally isolated from the structural elements and the pipe where the flow is measured. In between the two detectors, a heating element is positioned. Both detectors are connected to electrical wires through tiny conductors to minimize thermal loss through conduction (Fig. 11.4B). The sensor operates as follows. The first temperature detector R_0 measures the temperature of the flowing medium. The heater warms up the medium and the elevated temperature is measured by the second temperature detector R_s. In a still medium, heat would be dissipated from the heater through media to both detectors. In a medium with a zero flow, heat moves out from the heater mainly by thermal conduction and gravitational convection. Because the heater is positioned

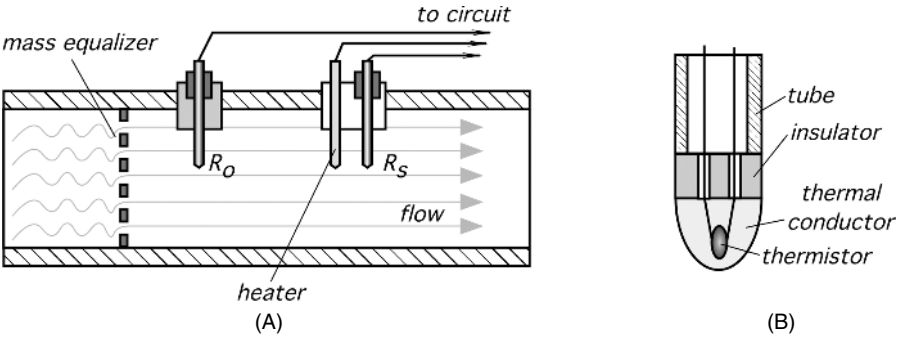


Fig. 11.4. Thermoanemometer. (A) a basic two-sensor design; (B) cross-sectional view of a temperature detector.

closer to the R_s detector, that detector will register a higher temperature. When the medium flows, heat dissipation increases due to forced convection. The higher the rate of flow, the higher the heat dissipation and the lower the temperature that will be registered by the R_s detector. Heat loss is measured and converted into the flow rate of the medium.

A fundamental relationship of thermoanemometry is based on King's law [2]

$$\Delta Q = kl\left(1+\sqrt{\frac{2\pi\rho c\, dv}{k}}\right)(T_s - T_0)\,, \tag{11.12}$$

where k and c are the thermal conductivity and specific heat of a medium at a given pressure, ρ is the density of the medium, l and d are the length and diameter of the sensor, respectively, T_s is the surface temperature of the sensor, T_0 is the temperature of the medium away from the sensor, and v is the velocity of the medium. Collis and Williams experimentally proved [3] that King's theoretical law needs some correction. For a cylindrical sensor with $l/d \gg 1$, a modified King's equation yields the velocity of the medium:

$$v = \frac{K}{\rho}\left(\frac{dQ}{dt}\frac{1}{T_s - T_0}\right)^{1.87}, \tag{11.13}$$

where K is the calibration constant. It follows from the above that to measure a flow, a temperature gradient between the sensor and the moving medium and the dissipated heat must be measured. Then, velocity of the fluid or gas becomes, although nonlinear, a quite definitive function of thermal loss (Fig. 11.5A).

To maintain the R_s detector at T_s and to assure a sufficient thermal gradient with respect to T_0, heat loss must be compensated for by supplying the appropriate power to the heating element. Also, we may consider a flow sensor without a separate heating element. In such a sensor, the R_s detector operates in a self-heating mode; that is, the electric current passing through its resistance generates enough Joule heat to elevate its temperature to T_s. At that temperature, the second detector has resistance R_s. Assuming that conductive heat loss through connecting wires and sensor's enveloping

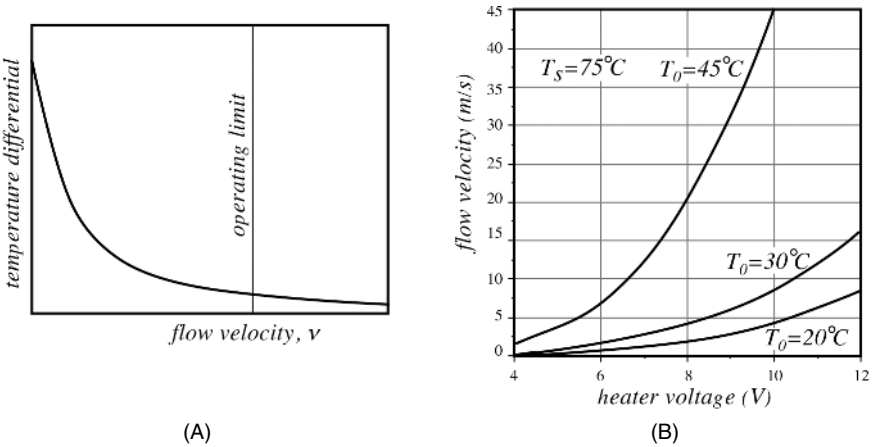


Fig. 11.5. Transfer function of a thermoanemometer (A) and calibration curves for a self-heating sensor in a thermoanemometer for three different levels of heat (B).

tube is negligibly small, the law of conservation of energy demands that electric power W be equal to thermal loss to flowing medium:

$$W = \frac{dQ}{dt}. \tag{11.14}$$

On the other hand, the electric power through a heating resistance is in a square relationship with the voltage e across the heating element:

$$W = \frac{e^2}{R_s}. \tag{11.15}$$

Equations (11.13)–(11.15) yield a relationship between the voltage across the self-heating detector and the velocity of flow:

$$v_{2a} = \frac{K}{\rho}\left(\frac{e^2}{R_s}\frac{1}{T_s - T_0}\right)^{1.87}. \tag{11.16}$$

Figure 11.5B shows an example of a calibrating curve for a flow sensor using a self-heating thermistor ($T_s = 75°C$) operating in air whose temperature varies from 20°C to 45°C. The thermistor temperature was maintained constant over an entire range of T_0 temperatures.[2] It should be emphasized, that T_s must always be selected higher than the highest temperature of the flowing medium.

Formula (11.13) suggests that two methods of measurement are possible. In the first method, the voltage and resistance of a heating element is maintained constant,

[2] This can be accomplished by using a self-balancing resistive bridge. See, for example, Ref. [4].

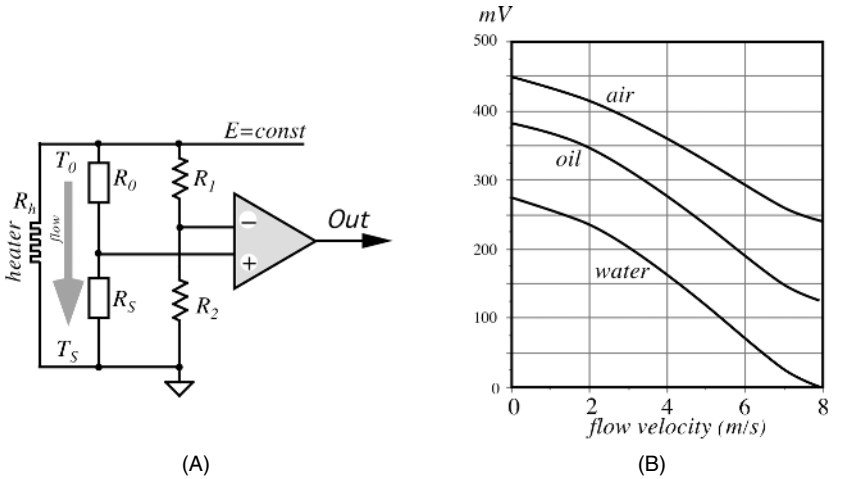


Fig. 11.6. Bridge circuit for a thermal flowmeter (A); sensor responses for different fluids (B).

whereas the temperature differential $(T_s - T_0)$ is used as the output signal. In the second method, the temperature differential is maintained constant by a control circuit which regulates the heater's voltage e. In the latter case, e is the output signal. This method is often preferable for use in the miniature sensors where self-heating temperature detectors are employed. A self-heating sensor [it can be either a resistive temperature detector (RTD) or thermistor] operates at high excitation currents. That current serves two purposes: It measures the resistance of a detector to determine its temperature and it provides Joule heat. Figure 11.6A shows that both temperature detectors (heated and reference) can be connected in a bridge circuit. At very low flow velocities, the bridge is imbalanced and the output signal is high. When the flow rate increases, the heated detector cools down and its temperature comes closer to that of a reference detector, lowering the output voltage. Figure 11.6B illustrates that the sensor's response is different for various fluids and gases. A sensor manufacturer usually provides calibration curves for any particular medium; however, whenever precision measurement is required, on-site calibration is recommended.

For accurate temperature measurements in a flowmeter, any type of temperature detector can be used: resistive, semiconductor, optical, and so forth (Chapter 17). Currently, however, the majority of manufacturers use resistive sensors. In industry and scientific measurements, RTDs are the prime choice, as they assure higher linearity, predictable response, and long-term stability over broader temperature ranges. In medicine, thermistors are often preferred because of their higher sensitivity. Whenever a resistive temperature sensor is employed, especially for a remote sensing, a four-wire measurement technique should be seriously considered. The technique is a solution for a problem arising from a finite resistance of connecting wires which may be a substantial source of error, especially with low-resistance-temperature sensors like RTDs. See Section 5.8.2 of Chapter 5 for the description of a four-wire method.

A sensor's design determines its operating limits. At a certain velocity, the molecules of a moving medium while passing near a heater do not have sufficient time to absorb enough thermal energy for developing a temperature differential between two detectors. Because the differential is in the denominator of Eq. (11.13), at high velocities computational error becomes unacceptably large and accuracy drops dramatically. The upper operating limits for the thermal transport sensors usually are determined experimentally. For instance, under normal atmospheric pressure and room temperature (about 20°C), the maximum air velocity that can be detected by a thermal transport sensor is in the range of 60 m/s (200 ft/s).

While designing thermal flow sensors, it is important to assure that the medium moves through the detectors without turbulence in a nonlaminar well-mixed flow. The sensor is often supplied with mixing grids or turbulence breakers which sometimes are called *mass equalizers* (Fig. 11.4A).

The pressure and temperature of a moving medium, especially of gases, make a strong contribution to the accuracy of a volume rate calculation. It is interesting to note that for the mass flow meters, pressure makes very little effect on the measurement as the increase in pressure results in a proportional increase in mass.

A data processing system for the thermal transport sensing must receive at least three variable input signals: a flowing medium temperature, a temperature differential, and a heating power signal. These signals are multiplexed, converted into digital form, and processed by a computer to calculate characteristics of flow. Data are usually displayed as velocity (m/s or ft/s), volume rate (m^3/s or ft^3/s), or mass rate (kg/s or lb/s).

Thermal transport flowmeters are far more sensitive than other types and have a broad dynamic range. They can be employed to measure very minute gas or liquid displacements as well as fast and strong currents. Major advantages of these sensors are the absence of moving components and an ability to measure very low flow rates. "Paddle wheel," hinged vane, and pressure differential sensors have low and inaccurate outputs at low rates. If a small-diameter tubing is required, as in automotive, aeronautic, medical, and biological applications, sensors with moving components become mechanically impractical. In these applications, thermal transport sensors are indispensable.

11.4 Ultrasonic Sensors

Flow can be measured by employing ultrasonic waves. The main idea behind the principle is the detection of frequency or phase shift caused by flowing medium. One possible implementation is based on the Doppler effect (see Section 6.2 of Chapter 6 for the description of the Doppler effect), whereas the other relies on the detection of the increase or decrease in effective ultrasound velocity in the medium. The effective velocity of sound in a moving medium is equal to the velocity of sound relative to the medium plus the velocity of the medium with respect to the source of the sound. Thus, a sound wave propagating upstream will have a smaller effective velocity, and the sound propagating downstream will have a higher effective velocity. Because the difference between the two velocities is exactly twice the velocity of the medium,

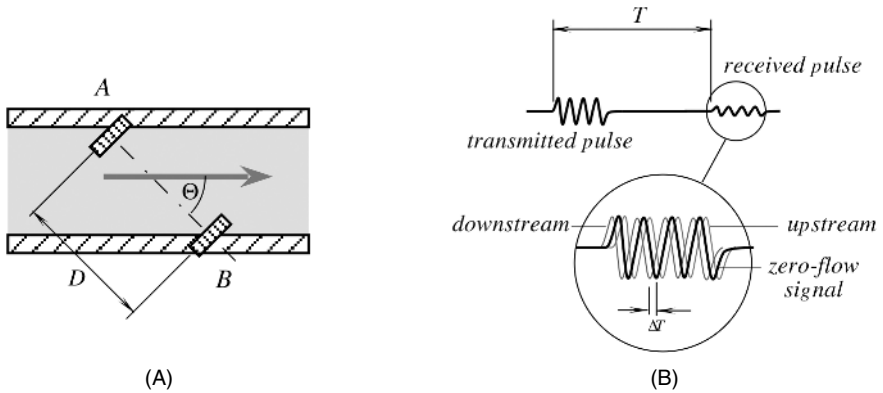


Fig. 11.7. Ultrasonic flowmeter. (A) position of transmitter–receiver crystals in the flow; (B) waveforms in the circuit.

measuring the upstream–downstream velocity difference allows us to determine the velocity of the flow.

Figure 11.7A shows two ultrasonic generators positioned on opposite sides of a tube of flow. Piezoelectric crystals are usually employed for that purpose. Each crystal can be used for either the generation of the ultrasonic waves (motor mode) or for receiving the ultrasonic waves (generator mode). In other words, the same crystal, when needed, acts as a "speaker" or a "microphone."

Two crystals are separated by distance D and positioned at angle Θ with respect to flow. Also, it is possible to place small crystals right inside the tube along the flow. That case corresponds to $\Theta = 0$. The transit time of sound between two transducers A and B can be found through the average fluid velocity v_c:

$$T = \frac{D}{c \pm v_c \cos \Theta}, \tag{11.17}$$

where c is the velocity of sound in the fluid. The plus and minus signs refer to the downstream and upstream directions, respectively. The velocity v_c is the flow velocity averaged along the path of the ultrasound. Gessner [4] has shown that for laminar flow $v_c = 4v_a/3$, and for turbulent flow, $v_c = 1.07v_a$, where v_a is the flow averaged over the cross-sectional area. By taking the difference between the downstream and upstream velocities, we find [5]

$$\Delta T = \frac{2Dv_c \cos \Theta}{c^2 + v_c \cos^2 \Theta} \approx \frac{2Dv_c \cos \Theta}{c^2}, \tag{11.18}$$

which is true for the most practical cases when $c \gg v_c \cos \Theta$. To improve the signal-to-noise ratio, the transit time is often measured for both upstream and downstream directions; that is, each piezoelectric crystal works as a transmitter at one time and as a receiver at the other time. This can be accomplished by a selector (Fig. 11.8) which is clocked by a relatively slow sampling rate (400 Hz in the example). The

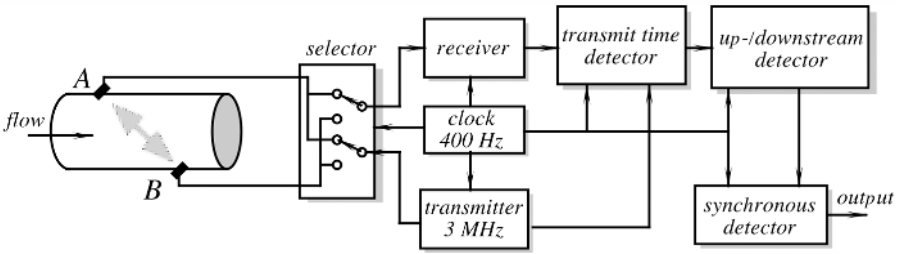


Fig. 11.8. Block diagram of an ultrasonic flowmeter with alternating transmitter and receiver.

sinusoidal ultrasonic waves (about 3 MHz) are transmitted as bursts with the same slow clock rate (400 Hz). A received sinusoidal burst is delayed from the transmitted one by time T, which is modulated by the flow (Fig. 11.7B). This time is detected by a transit-time detector; then, the time difference in both directions is recovered by a synchronous detector. Such a system can achieve quite good accuracy, with a zero drift as small as 5×11^{-3} m/s^2 over the 4-h period.

An alternative way of measuring flow with ultrasonic sensors is to detect a phase difference in transmitted and received pulses in the upstream and downstream directions. The phase differential can be derived from Eq. (11.18):

$$\Delta f = \frac{4\pi f D v_c \cos\Theta}{c^2}, \tag{11.19}$$

where f is the ultrasonic frequency. It is clear that the sensitivity is better with the increase in the frequency; however, at higher frequencies, one should expect stronger sound attenuation in the system, which may cause a reduction in the signal-to-noise ratio.

For the Doppler flow measurements, continuous ultrasonic waves can be used. Figure 11.9 shows a flowmeter with a transmitter–receiver assembly positioned inside the flowing stream. As in a Doppler radio receiver, transmitted and received frequen-

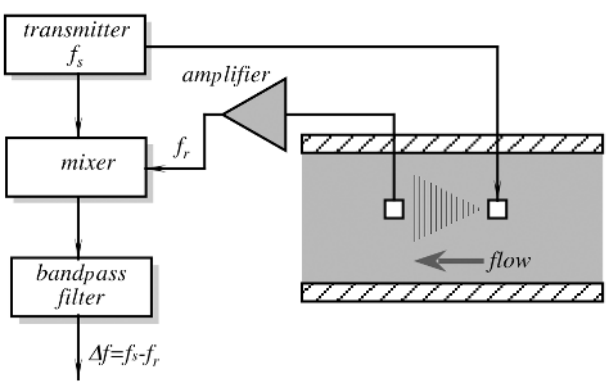


Fig. 11.9. Ultrasonic Doppler flowmeter.

cies are mixed in a nonlinear circuit (a mixer). The output low-frequency differential harmonics are selected by a bandpass filter. That differential is defined as

$$\Delta f = f_s - f_r \approx \pm \frac{2 f_s v}{c}, \tag{11.20}$$

where f_s and f_r are the frequencies in the transmitting and receiving crystals, respectively, and the plus/minus signs indicate different directions of flow. An important conclusion from the above equation is that the differential frequency is directly proportional to the flow velocity. Obviously, the crystals must have much smaller sizes than the clearance of the tube of flow. Hence, the measured velocity is not the average but rather a localized velocity of flow. In practical systems, it is desirable to calibrate ultrasonic sensors with actual fluids over the useful temperature range, so that contribution of a fluid viscosity is taken into account.

An ultrasonic piezoelectric sensors/transducer can be fabricated of small ceramic disks encapsulated into a flowmeter body. The surface of the crystal can be protected by a suitable material, (e.g., silicone rubber). An obvious advantage of an ultrasonic sensor is in its ability to measure flow without a direct contact with the fluid.

11.5 Electromagnetic Sensors

The electromagnetic flow sensors are useful for measuring the movement of conductive liquids. The operating principle is based on the discovery of Faraday and Henry (see Section 3.4 of Chapter 3) of the electromagnetic induction. When a conductive media (wire, for instance) or for this particular purpose, flowing conductive liquid crosses the magnetic flux lines, the electromotive force (e.m.f.) is generated in the conductor. The value of the e.m.f. is proportional to velocity of moving conductor [Eq. (3.37) of Chapter 3]. Figure 11.10 illustrates a tube of flow positioned into magnetic field **B**. There are two electrodes incorporated into a tube to pick up the e.m.f.

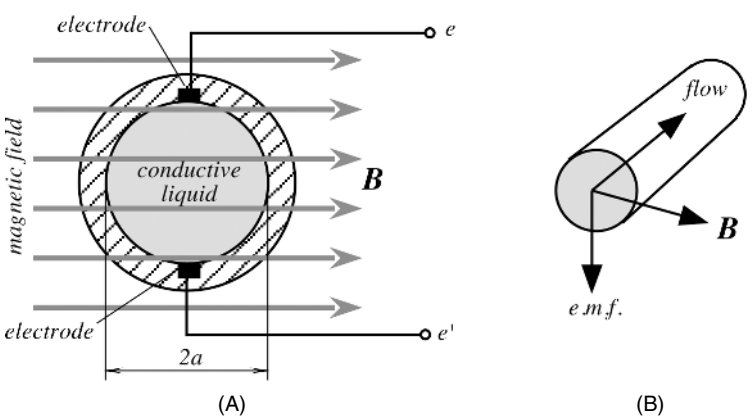


Fig. 11.10. Principle of an electromagnetic flowmeter: (A) position of electrodes is perpendicular to the magnetic field; (B) relationships between flow and electrical and magnetic vectors.

induced in the liquid. The magnitude of the e.m.f. is defined by

$$v = e - e' = 2a\mathbf{B}v, \tag{11.21}$$

where a is the radius of the tube of flow and v is the velocity of flow.

By solving Maxwell's equations, it can be shown that for a typical case when the fluid velocity is nonuniform within the cross-sectional area but remains symmetrical about the tube axis (axisymmetrical), the e.m.f generated is the same as that given by Eq. (11.21), except that v is replaced by the average velocity, v_a [Eq. (11.3)]:

$$v_a = \frac{1}{\pi a^2} \int_0^a 2\pi v r \, dr, \tag{11.22}$$

where r is the distance from the center of the tube. Equation (11.21) can be expressed in terms of the volumetric flow rate:

$$v = \frac{2\Lambda\mathbf{B}}{\pi a}. \tag{11.23}$$

It follows from Eq. (11.23) that the voltage registered across the pickup electrodes is independent of the flow profile or fluid conductivity. For a given tube geometry and the magnetic flux, it depends only on the instantaneous volumetric flow rate.

There are two general methods of inducing voltage in the pickup electrodes. The first is a dc method where the magnetic flux density is constant and induced voltage is a dc or slow-changing signal. One problem associated with this method is a polarization of the electrodes due to small but unidirectional current passing through their surface. The other problem is a low-frequency noise, which makes it difficult to detect small flow rates.

Another and far better method of excitation is with an alternating magnetic field, which causes the appearance of an ac voltage across the electrodes (Fig. 11.11). Naturally, the frequency of the magnetic field should meet a condition of the Nyquist rate; that is, it must be at least two times higher than the highest frequency of flow-rate spectrum variations. In practice, the excitation frequency is selected in the range between 100 and 1000 Hz.

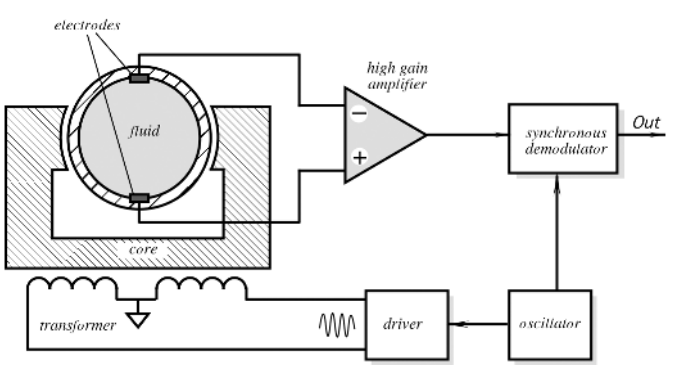


Fig. 11.11. Electromagnetic flowmeter with synchronous (phase-sensitive) demodulator.

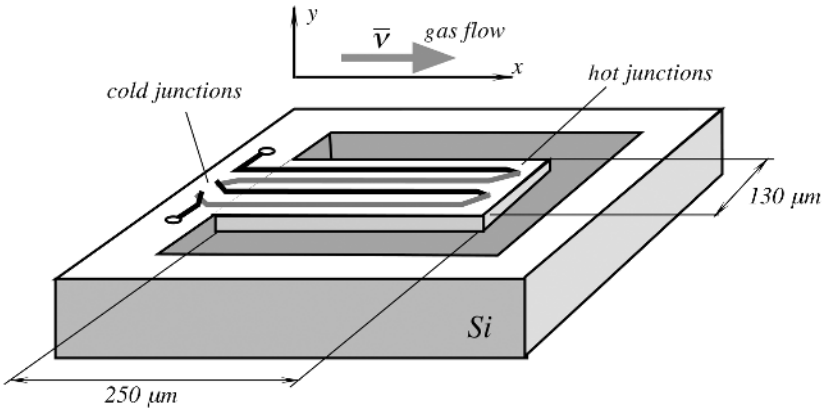


Fig. 11.12. Micromachined gas flow sensor.

11.6 Microflow Sensors

In some applications, such as process control in precise semiconductor manufacturing, chemical and pharmaceutical industries, and biomedical engineering, miniaturized gas flow sensors are encountered with increasing frequency. Most of them operate on the method of thermal transport (see Section 11.3) and are fabricated from a silicon crystal by using micromachining technology. Many of the microflow sensors use a thermopile as a temperature sensor [6]; however, the thermoelectric coefficient [Eq. (3.91) of Chapter 3] of standard elements used in the integrated circuit (IC) processing (silicon and aluminum) is smaller than that of conventional thermocouples by factors ranging from 10 to 100. Thus, a resulting output signal may be very small, which requires amplification by amplifiers integrated directly into the sensor.

A cantilever design of a microflow sensor is shown in Fig. 11.12. The thickness of the cantilever may be as low as 2 μm. It is fabricated in the form of a sandwich consisting of layers of field oxide, chemical vapor deposition (CVD) oxide, and nitrate [7]. The cantilever sensor is heated by an imbedded resistor with a rate of 26 K/mW of applied electric power, and a typical transfer function of the flow sensor has a negative slope of about 4 mV/m/s.

The heat is removed from the sensor by three means: conductance L_b through the cantilever beam, gas flow $h(v)$, and thermal radiation, which is governed by the Stefan–Boltzmann law:

$$P = L_b(T_s - T_b) + h(v)(T_s - T_b) + a\sigma\varepsilon(T_s^4 - T_b^4), \tag{11.24}$$

where σ is the Stefan–Boltzmann constant, a is the area along which the beam-to-gas heat transfer occurs, ε is surface emissivity, and v is the gas velocity. From the principles of energy and particle conservation, we deduce a generalized heat-transport equation governing the temperature distribution $T(x, y)$ in the gas flowing near the

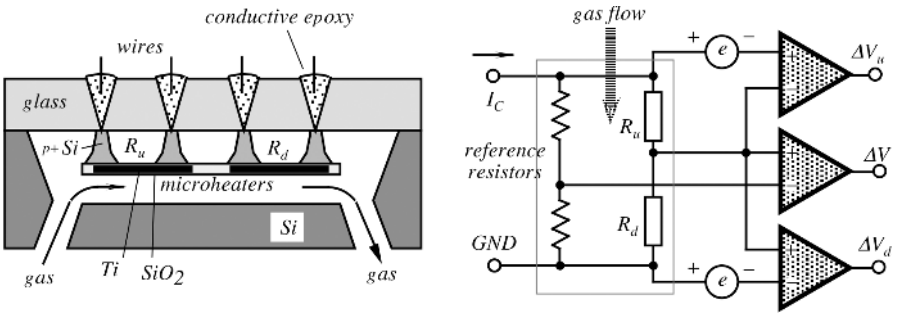


Fig. 11.13. Gas microflow sensor with self-heating titanium resistors: (A) sensor design; (B) interface circuit: R_u and R_d are resistances of the upstream and downstream heaters, respectively. (Adapted from Ref. [7].)

sensor's surface:

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = \frac{vnc_p}{k_g}\frac{\partial T}{\partial x} \quad \text{for } y > 0, \tag{11.25}$$

where n is the gas density, c_p is the molecular gas capacity, and k_g is the thermal conductivity of gas. It can be shown that the solution of this equation for the boundary condition of a vanishing thermal gradient far off the surface is [7]

$$\Delta V = B\left(\frac{1}{\sqrt{\mu^2+1}} - 1\right), \tag{11.26}$$

where V is the input voltage, B is a constant, and $\mu = Lvnc_p/2\pi k_g$, and L is the gas sensor contact length. This solution coincides very well with the experimental data.

Another design of a thermal transport microsensor is shown in Fig. 11.13A [8] where titanium films having a thickness of 0.1 μm serve as both the temperature sensors and the heaters. The films are sandwiched between two layers of SiO_2. Titanium was used because of its high TCR (temperature coefficient of resistance) and excellent adhesion to SiO_2. Two microheaters are suspended with four silicon girders at a distance of 20 μm from one another. The Ti film resistance is about 2 kΩ. Figure 11.13B shows a simplified circuit diagram for the sensor, which exhibits an almost linear relationship between the flow and output voltage ΔV.

A microflow sensor can be constructed by utilizing a capacitive pressure sensor [9] as shown in Fig. 11.14. An operating principle of the sensor is based on a pressure gradient technique as described in Section 11.2. The sensor was fabricated using silicon micromachining and defused boron etch-stops to define the structure. The gas enters the sensor's housing at pressure P_1 through the inlet, and the same pressure is established around the silicon plate, including the outer side of the etched membrane. The gas flows into the microsensor's cavity through a narrow channel having a relatively high-pressure resistance. As a result, pressure P_2 inside the cavity

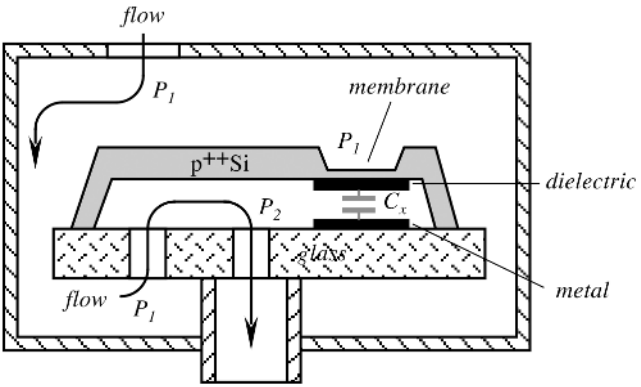


Fig. 11.14. Structure of a gas microflow sensor utilizing capacitive pressure sensor. (Adapted from Ref. [9].)

is lower than P_1, thus creating a pressure differential across the membrane. Therefore, the flow rate can be calculated from Eq. (11.10).

The pressure differential is measured by a capacitive pressure sensor, which is composed of a thin, stress-compensated, p^{++} boron-doped silicon membrane suspended above a metal plate. The pressure differential changes the capacitance C_x between the metal plate and the silicon structure with a resolution of 1 mTorr/fF, with a full pressure of about 4 torr. The overall resolution of the sensor is near 14–15 bits and the accuracy of pressure measurement about 9–10 bits. At approximately twice the full-scale pressure differential, the membrane touches the metal plate; hence, a dielectric layer is required to prevent an electric short, and the substrate glass plate protects the membrane from rupturing. A capacitance measurement circuit (see Fig. 5.32 of Chapter 5) is integrated with the silicon plate using standard CMOS technology.

11.7 Breeze Sensor

In some applications, it is desirable just to merely detect a change in air (or any other gas for that matter) movement, rather than to measure its flow rate quantitatively. This task can be accomplished by a breeze sensor, which produces an output transient whenever the velocity of the gas flow happens to change. One example of such a device is a piezoelectric breeze sensor produced by Nippon Ceramic of Japan. The sensor contains a pair of the piezoelectric (or pyroelectric) elements,[3] where one is exposed to ambient air and the other is protected by the encapsulating resin coating. Two sensors are required for a differential compensation of variations in ambient

[3] In this sensor, the crystalline element which is poled during the manufacturing process is the same as used in piezoelectric or pyroelectric sensors. However, the operating principle of the breeze sensor is neither related to mechanical stress nor heat flow. Nevertheless, for simplicity of the description, we will use the term *piezoelectric*.

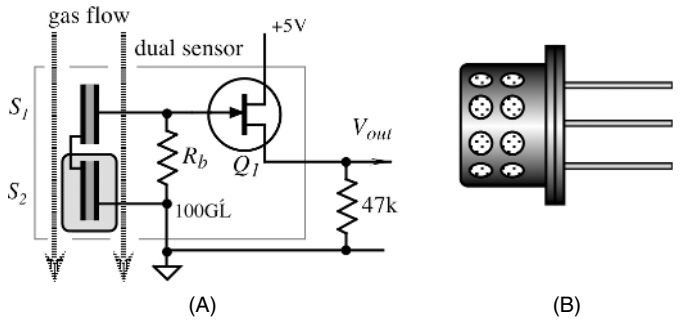


Fig. 11.15. Piezoelectric breeze sensor. (A) a circuit diagram; (B) a packaging in a TO-5 can.

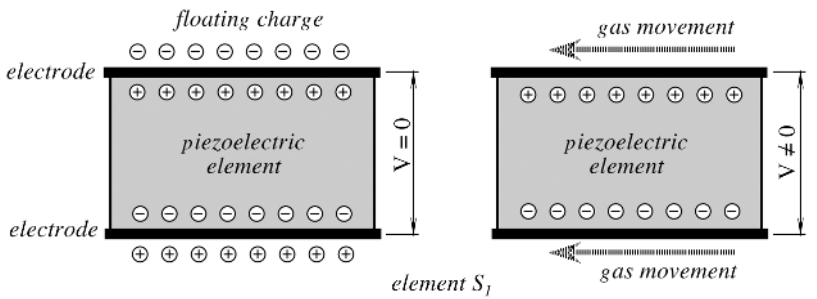


Fig. 11.16. In a breeze sensor, gas movement strips off electric charges from the surface of a piezoelectric element.

temperature. The elements are connected in a series-opposed circuit; that is, whenever both of them generate the same electric charge, the resulting voltage across the bias resistor R_b (Fig. 11.15A) is essentially zero. Both elements, the bias resistor and the JFET voltage follower, are encapsulated into a TO-5 metal housing with vents for exposing the S_1 element to the gas movement (Fig. 11.15B).

The operating principle of the sensor is illustrated in Fig. 11.16. When airflow is either absent or is very steady, the charge across the piezoelectric element is balanced. Element internal electric dipoles, which are oriented during the poling process (Section 3.6 of Chapter 3), are balanced by both the free carriers inside the material and the charged floating air molecules at the element's surface. As a result, voltage across the piezoelectric elements S_1 and S_2 is zero, which results in baseline output voltage V_{out}. When the gas flow across both S_1 surfaces changes (S_2 surfaces are protected by resin), moving gas molecules strip off the floating charges from the element. This results in the appearance of voltage across the element's electrodes, because the internally poled dipoles are no longer balanced by the outside floating charges. The voltage is repeated by the JFET follower, which serves as an impedance converter, and appears as a transient in the output terminal.

11.8 Coriolis Mass Flow Sensors

Coriolis flowmeters measure flow of mass directly, as opposed to those that measure velocity or volume [10]. Coriolis flowmeters are virtually unaffected by the fluid pressure, temperature, viscosity, and density. As a result, Coriolis meters can be used without recalibration and without compensating for parameters specific to a particular type of fluid. Although these meters were used mainly for liquids when they were first introduced, they have recently become adaptable for the gas applications.

Coriolis flowmeters are named after Gaspard G. Coriolis (1792–1843), a French civil engineer and physicist. A Coriolis sensor typically consists of one or two vibrating tubes with an inlet and an outlet. A typical material for the tube is stainless steel. It is critical for meter accuracy to prevent any mechanical or chemical attack of the tube or its lining by the flowing fluid. Some tubes are U-shaped but a wide variety of shapes have been also employed. The thinner tubes are used for gas, whereas thicker tubes are more appropriate for liquids. The Coriolis tube is set to vibration by an auxiliary electromechanical drive system.

Fluid enters the meter in the inlet. A mass flow is determined based on the action of the fluid on the vibrating tubes. As fluid moves from the inlet to outlet, it develops different forces depending on its acceleration that is the result of the tube vibration.

The Coriolis force induced by the flow is described by

$$F = 2m\omega v \tag{11.27}$$

where m is the mass, ω is the rotating circular frequency, and **v** is the vector of the average fluid velocity. As a result of these forces, the tube takes on a twisting motion as it passes though the vibrating cycle. The amount of twist is directly proportional to the mass flow through the tube. Figure 11.17A shows the Coriolis flow tube in a no-flow situation, and Fig. 11.17B shows Coriolis tube with the flow.

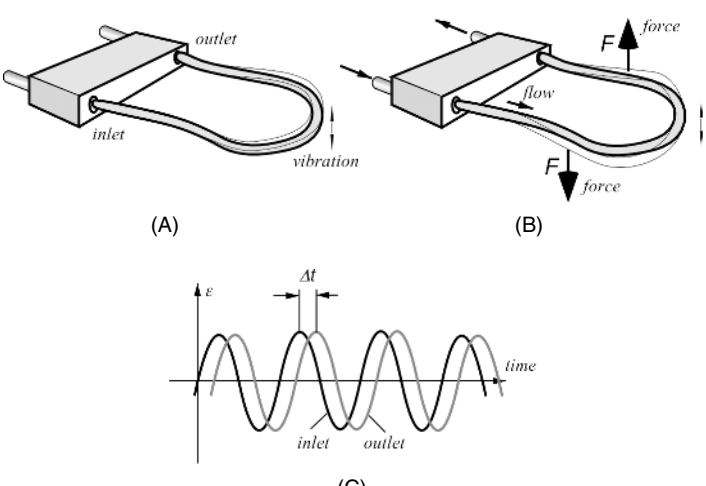


Fig. 11.17. Coriolis tube with no flow (A); twist of the tube with flow (B); vibrating phase shift resulting from Coriolis forces (C).

With a no-flow state, the tube vibrates identically at its inlet and outlet sides with the sine-wave motions with the zero phase shift between them. During flow, the tube twists in response to the flow, and the inlet and outlet sides vibrate differently with a phase shift between them (Fig. 11.17C). The main disadvantage of the Coriolis sensor its relatively high initial cost. However, the versatility of Coriolis sensors in handling multiple fluids makes them very useful for plants where the flow of multiple fluid types must be measured. There are also an increasing number of the gas applications for the Coriolis meters.

11.9 Drag Force Flow Sensors

When fluid motion is sporadic, multidirectional, and turbulent, a drag force flow sensor may be quite efficient. Application of such flowmeters include environmental monitoring, meteorology, hydrology, and maritime studies to measure the speed of air or water flow and turbulence close to surface [11]. In the *flowmeter*, a solid object known as a *drag element* or *target* is exposed to the flow of fluid. The force exerted by the fluid on the drag element is measured and converted to a value for speed of flow. An important advantage of the drag sensor is that it can be made to generate a measurement of flow in two dimensions, or even in three dimensions, as well as of flow speed. To implement this feature, the drag element must be symmetrical in the appropriate number of dimensions. These flowmeters have been used by industry, utilities, aerospace, and research laboratories to measure the flow of unidirectional and bidirectional liquids (including cryogenic), gases, and steam (both saturated and superheated) for almost half a century.

The operation of the sensor is based on strain measurement of deformation of an elastic rubber cantilever, to which a force is applied by a spherical symmetrical drag element (Fig. 11.18). However, an ideal drag element is a flat disk [12], because this configuration gives a drag coefficient independent of the flow rate. Using a spherical drag element, which departs from the ideal of a flat disk, the drag coefficient may vary with flow speed, and, therefore, the gauge must be calibrated and optimized for the conditions of intended use. The strain measurement can be performed with strain gauges that should be physically protected from interaction with moving fluids.

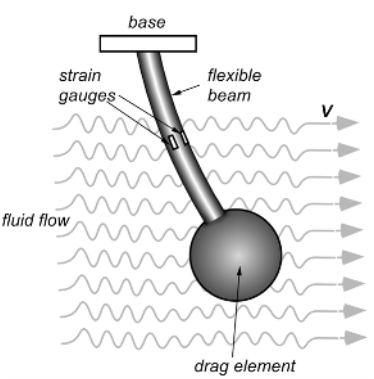


Fig. 11.18. Drag force sensor.

The drag force F, exerted by incompressible fluid on a solid object exposed to it is given by the drag *equation*:

$$F_D = C_D \rho A V^2, \tag{11.28}$$

where ρ is the fluid density, V is the fluid velocity at the point of measurement, A is the projected area of the body normal to the flow, and C_D is the overall drag coefficient. C_D is a dimensionless factor whose magnitude depends primarily on the physical shape of the object and its orientation relative to the fluid stream. If mass of the supporting beam is ignored, the developed strain is

$$\varepsilon = \frac{3 C_D \rho A V^2 (L - x)}{E a^2 b}, \tag{11.29}$$

where L is the beam length, x is the point coordinate on the beam where the strain gauges are located, E is Young's modulus of elasticity, and a and b are the geometry target factors. It is seen that the strain in a beam is a square-law function of the fluid speed.

References

1. Benedict, R. P. *Fundamentals of Temperature, Pressure, and Flow Measurements*, 3rd ed. John Wiley & Sons, New York, 1984.
2. King, L.V. On the convention of heat from small cylinders in a stream of fluid. *Phil. Trans. Roy. Soc.* A214, 373, 1914.
3. Collis, D. C. and Williams, M. J. Two-dimensional convection from heated wires at low Reynolds' numbers. *J. Fluid Mech.* 6, 357, 1959.
4. Gessner, U. The performance of the ultrasonic flowmeter in complex velocity profiles. *IEEE Trans. Bio-Med. Eng.* MBE-16, 139–142, 1969.
5. Cobbold, R.S.C. *Transducers for Biomedical Measurements*. John Wiley & Sons, New York, 1974.
6. Van Herwaarden, A.W. and Sarro, P.M. Thermal sensors based on the Seebeck effect. *Sensors Actuators* 10, 321–346, 1986.
7. Wachutka, G., Lenggenhager, R., Moser, D, and Baltes, H. Analytical 2D-model of CMOS micromachined gas flow sensors. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*. IEEE, New York, 1991.
8. Esashi, M. Micro flow sensor and integrated magnetic oxygen sensor using it. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*. IEEE, New York, 1991.
9. Cho, S.T. and Wise, K.D. A high performance microflowmeter with built-in self test. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*, IEEE, New York, 1991, pp. 400–403.
10. Yoder, J. Coriolis Effect Mass Flowmeters. In: *Mechanical Variables Measurement*, J. Webster, ed. CRC Press, Boca Raton, FL, 2000.

11. Philip-Chandy, R., Morgan,R and Scully, P.J. Drag force flowmeters. In: *Mechanical Variables Measurement*, J. Webster, ed. CRC Press, Boca Raton, FL, 2000.
12. Clarke T. Design and operation of target flowmeters. In: *Encyclopedia of Fluid Mechanics, Vol. 1*. Gulf Publishing, Houston, TX, 1986.

This page intentionally left blank

12

Acoustic Sensors

"Your ears will always lead you right,
but you must know why."

—Anton von Webern

The fundamentals of acoustics are given in Section 3.10 of Chapter 3. Here, we will discuss the acoustic sensors for various frequency ranges. The audible range sensors are generally called the *microphones*; however, the name is often used even for the ultrasonic and infrasonic waves. In essence, a microphone is a pressure transducer adapted for the transduction of sound waves over a broad spectral range which generally excludes very low frequencies below a few hertz. The microphones differ by their sensitivity, directional characteristics, frequency bandwidth, dynamic range, sizes, and so forth. Also, their designs are quite different depending on the media from which sound waves are sensed. For example, for the perception of air waves or vibrations in solids, the sensor is called a *microphone*, whereas for the operation in liquids, it is called a *hydrophone* (even if the liquid is not water—from the Greek name of mythological water serpent Hydra). The main difference between a pressure sensor and an acoustic sensor is that latter does not need to measure constant or very slow-changing pressures. Its operating frequency range usually starts at several hertz (or as low as tens of millihertz for some applications), and the upper operating frequency limit is quite high—up to several megahertz for the ultrasonic applications and even gigahertz in the surface acoustic-wave device.

Because acoustic waves are mechanical pressure waves, any microphone or hydrophone has the same basic structure as a pressure sensor: it is composed of a moving diaphragm and a displacement transducer which converts the diaphragm's deflections into an electrical signal; that is, all microphones or hydrophones differ by the design of these two essential components. Also, they may include some additional parts such as mufflers, focusing reflectors or lenses, and so forth; however, in this chapter, we will review only the sensing parts of some of the most interesting, from our point of view, acoustic sensors.

12.1 Resistive Microphones

In the past, resistive pressure converters were used quite extensively in microphones. The converter consisted of a semiconductive powder (usually graphite) whose bulk resistivity was sensitive to pressure. Currently, we would say that the powder possessed piezoresistive properties. However, these early devices had quite a limited dynamic range, poor frequency response, and a high noise floor. Presently, the same piezoresistive principle can be employed in the micromachined sensors, where stress-sensitive resistors are the integral parts of a silicon diaphragm (Section 10.5 of Chapter 10).

12.2 Condenser Microphones

If a parallel-plate capacitor is given an electric charge q, the voltage across its plates is governed by Eq. (3.19 of Chapter 3). On the other hand, according to Eq. (3.20 of Chapter 3) the capacitance depends on distance d between the plates. Thus, solving these two equations for voltage, we arrive at

$$V = q\frac{d}{\varepsilon_0 A}, \tag{12.1}$$

where $\varepsilon_0 = 8.8542 \times 10^{-12}$ C^2/N m^2 is the permitivity constant (Section 3.1 of Chapter 3). Equation (12.1) is the basis for operation of the *condenser* microphones, which is another way to say "capacitive" microphones. Thus, a capacitive microphone linearly converts a distance between the plates into electrical voltage which can be further amplified. The device essentially requires a source of an electric charge q whose magnitude directly determines the microphone sensitivity. The charge can be provided either from an external power supply having a voltage in the range from 20 to 200 V or from an internal source capable of producing such a charge. This is accomplished by a built-in electret layer which is a polarized dielectric crystal.

Presently, many condenser microphones are fabricated with silicon diaphragms, which serve two purposes: to convert acoustic pressure into displacement and to act as a moving plate of a capacitor. Some promising designs are described in Refs. [1–3]. To achieve high sensitivity, a bias voltage should be as large as possible, resulting in a large static deflection of the diaphragm, which may result in reduced shock resistivity and lower dynamic range. In addition, if the air gap between the diaphragm and the backplate is very small, the acoustic resistance of the air gap will reduce the mechanical sensitivity of the microphone at higher frequencies. For instance, at an air gap of 2 μm, an upper cutoff frequency of only 2 kHz has been measured [1].

One way to improve the characteristics of a condenser microphone is to use a mechanical feedback from the output of the amplifier to the diaphragm [4]. Figure 12.1A shows a circuit diagram and Fig. 12.1B is a drawing of interdigitized electrodes of the microphone. The electrodes serve different purposes: One is for the conversion of a diaphragm displacement into voltage at the input of the amplifier A_1 and the other electrode is for converting feedback voltage V_a into a mechanical deflection by means of electrostatic force. The mechanical feedback clearly improves the linearity and the frequency range of the microphone; however, it significantly reduces the deflection, which results in a lower sensitivity.

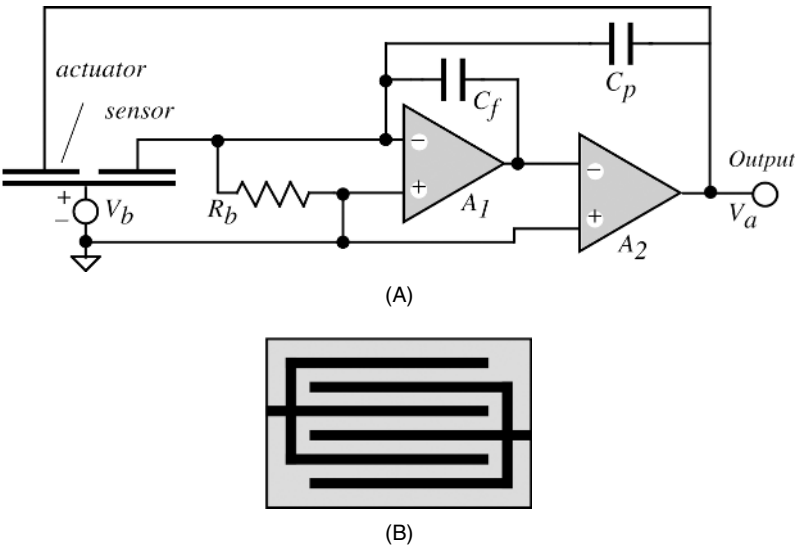


Fig. 12.1. Condenser microphone with a mechanical feedback: (A) a circuit diagram; (B) interdigitized electrodes on the diaphragm. (Adapted from Ref. [4].)

For further reading on condenser microphones an excellent book edited by Wong and Embleton is recommended [5].

12.3 Fiber-Optic Microphone

Direct acoustic measurements in hostile environments, such as in turbojets or rocket engines, require sensors which can withstand high heat and strong vibrations. The acoustic measurements under such hard conditions are required for computational fluid dynamics (CFD) code validation, structural acoustic tests, and jet noise abatement. For such applications, a fiber-optic interferometric microphone can be quite suitable. One such design [6] is composed of a single-mode temperature insensitive Michelson interferometer and a reflective plate diaphragm. The interferometer monitors the plate deflection, which is directly related to the acoustic pressure. The sensor is water cooled to provide thermal protection for the optical materials and to stabilize the mechanical properties of the diaphragm.

To provide an effect of interference between the incoming and outgoing light beams, two fibers are fused together and cleaved at the minimum tapered region (Fig. 12.2). The fibers are incorporated into a stainless-steel tube, which is water cooled. The internal space in the tube is filled with epoxy, and the end of the tube is polished until the optical fibers are observed. Next, aluminum is selectively deposited at one of the fused fiber core ends to make its surface mirror reflective. This fiber serves as a reference arm of the microphone. The other fiber core is left open and serves as the sensing arm. Temperature insensitivity is obtained by the close proximity of the reference and sensing arms of the assembly.

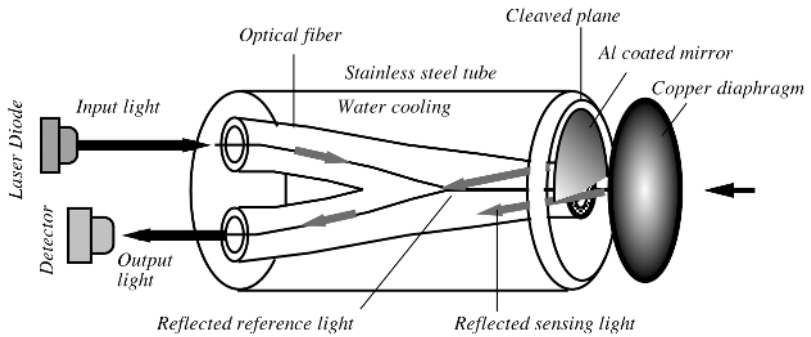


Fig. 12.2. Fiber-optic interferometric microphone. Movement of the copper diaphragm is converted into light intensity in the detector.

Light from a laser source (a laser diode operating near 1.3 μm wavelength) enters one of the cores and propagates toward the fused end, where it is coupled to the other fiber core. When reaching the end of the core, light in the reference core is reflected from the aluminum mirror toward the input and output sides of the sensor. The portion of light which goes toward the input is lost and has no effect on the measurement, whereas the portion which goes to the output strikes the detector's surface. That portion of light which travels to the right in the sensing core, exits the fiber, and strikes the copper diaphragm. Part of the light is reflected from the diaphragm back toward the sensing fiber and propagates to the output end, along with the reference light. Depending on the position of the diaphragm, the phase of the reflected light will vary, thus becoming different from the phase of the reference light.

While traveling together to the output detector, the reference and sensing lights interfere with one another, resulting in the light-intensity modulation. Therefore, the microphone converts the diaphragm displacement into a light intensity. Theoretically, the signal-to-noise ratio in such a sensor is obtainable on the order of 70–80 dB, thus resulting in an average minimum detectable diaphragm displacement of 1 Å (10^{-10} m).

Figure 12.3 shows a typical plot of the optical intensity in the detector versus the phase for the interference patterns. To assure a linear transfer function, the operating

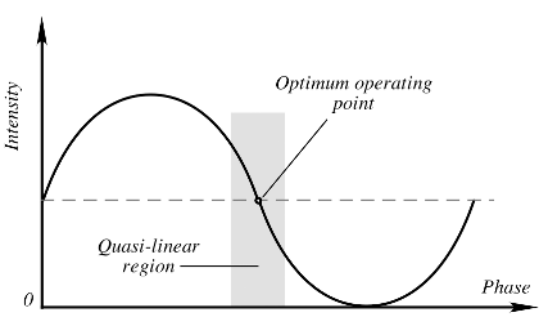


Fig. 12.3. Intensity plot as function of a reflected light phase.

point should be selected near the middle of the intensity, where the slope is the highest and the linearity is the best. The slope and the operating point may be changed by adjusting the wavelength of the laser diode. It is important for the deflection to stay within one-quarter of the operating wavelength to maintain a proportional input.

The diaphragm is fabricated from a 0.05-mm foil with a 1.25-mm diameter. Copper is selected for the diaphragm because of its good thermal conductivity and relatively low modulus of elasticity. The latter feature allows us to use a thicker diaphragm, which provides better heat removal while maintaining a usable natural frequency and deflection. A pressure of 1.4 kPa produces a maximum center deflection of 39 nm (390 AA), which is well within a one-quarter of the operating wavelength (1300 nm). The maximum acoustic frequency which can be transferred with the optical microphone is limited to about 100 kHz, which is well above the desired working range needed for the structural acoustic testing.

12.4 Piezoelectric Microphones

The piezoelectric effect can be used for the design of simple microphones. A piezoelectric crystal is a direct converter of a mechanical stress into an electric charge. The most frequently used material for the sensor is a piezoelectric ceramic, which can operate up to a very high frequency limit. This is the reason why piezoelectric sensors are used for the transduction of ultrasonic waves (Section 7.6 of Chapter 7). Still, even for the audible range, the piezoelectric microphones are used quite extensively. Typical applications are voice-activated devices and blood pressure measurement apparatuses where the arterial Korotkoff sounds have to be detected. For such acoustically nondemanding applications, the piezoelectric microphone design is quite simple (Fig. 12.4). It consists of a piezoelectric ceramic disk with two electrodes deposited on each side. The electrodes are connected to wires either by electrically conductive epoxy or by soldering. Because the output impedance of such a microphone is very large, a high-input-impedance amplifier is required.

Piezoelectric films [polyvinylidene fluoride (PVDF) and copolymers] were used for many years as very efficient acoustic pickups in musical instruments [7]. One of the first applications for piezoelectric film was as an acoustic pickup for a violin. Later, the

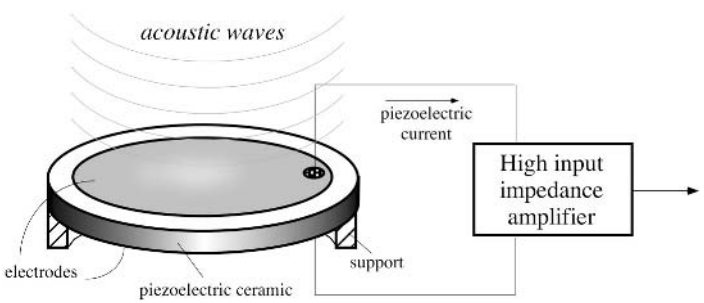


Fig. 12.4. Piezoelectric microphone.

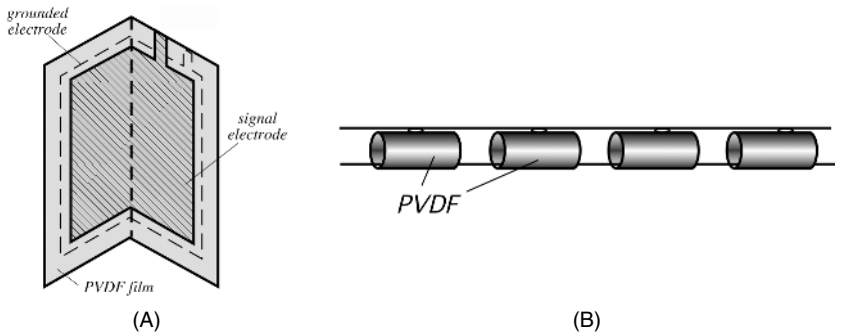


Fig. 12.5. Foldover piezoelectric acoustic pickup (A) and arrangement of a piezoelectric film hydrophone (B).

film was introduced for a line of acoustic guitars as a saddle-mounted bridge pickup, mounted in the bridge. The very high fidelity of the pickup led the way to a family of vibration-sensing and accelerometer applications: in one guitar pickup, a thick-film, compressive (under the saddle) design; another is a low-cost accelerometer, and another is an after-market pickup design that is taped to the instrument. Because of the low Q of the material, these transducers do not have the self-resonance of hard ceramic pickups. Shielding can be achieved by a foldover design as shown in Fig. 12.5A. The sensing side is the slightly narrower electrode on the inside of the fold. The foldover technique provides a more sensitive pickup than alternative shielding methods because the shield is formed by one of the electrodes. For application in water, the film can be rolled in tubes, and many of such tubes can be connected in parallel (Fig. 12.5B).

12.5 Electret Microphones

An electret is a close relative of piezoelectric and pyroelectric materials. In effect, they are all electrets with either enhanced piezoelectric or pyroelectric properties. An electret is a permanently electrically polarized crystalline dielectric material. The first application of electrets to microphones and earphones where described in 1928 [8]. An electret microphone is an electrostatic transducer consisting of a metallized electret and backplate separated from the diaphragm by an air gap (Fig. 12.6).

The upper metallization and a metal backplate are connected through a resistor R's voltage V across which it can be amplified and used as an output signal. Because the electret is a permanently electrically polarized dielectric, the charge density σ_1 on its surface is constant and sets an electric field E_1 in the air gap. When an acoustic wave impinges on the diaphragm, the latter deflects downward, reducing the air gap thickness s_1 for a value of Δs. Under open-circuit conditions, the amplitude of a variable portion of the output voltage becomes

$$V = \frac{s\,\Delta s}{\varepsilon_0(s + \varepsilon s_1)}. \tag{12.2}$$

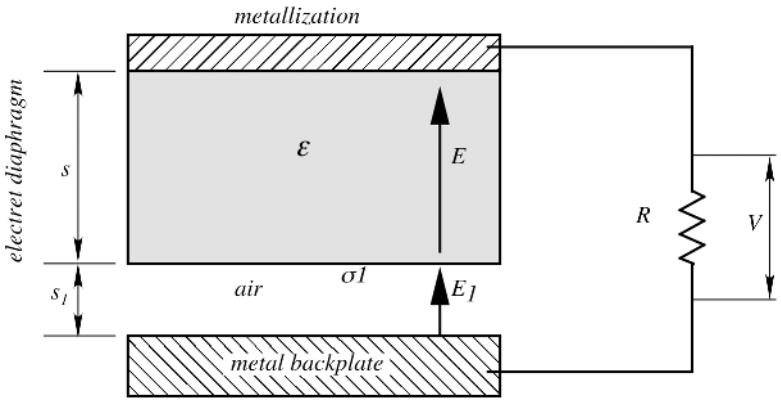


Fig. 12.6. General structure of an electret microphone. The thicknesses of layers are exaggerated for clarity. (After Ref. [9].)

Thus, the deflected diaphragm generates voltage across the electrodes. That voltage is in phase with the diaphragm deflection. If the sensor has a capacitance C, Eq. (12.2) should be written

$$V = \frac{s\Delta s}{\varepsilon_0(s + \varepsilon s_1)} \frac{2\pi f RC}{\sqrt{1 + (2\pi f RC)^2}}, \tag{12.3}$$

where f is the frequency of sonic waves.

If the restoring forces are due to the elasticity of the air cavities behind the diaphragm (effective thickness is s_0) and the tension T of the membrane, its displacement Δs to a sound pressure Δp assuming negligible losses is given by [10]

$$\Delta s = \frac{\Delta p}{(\gamma p_0/s_0) + (8\pi T/A)}, \tag{12.4}$$

where γ is the specific heat ratio, p_0 is the atmospheric pressure, and A is the membrane area. If we define the electret microphone sensitivity as $\delta_m = \Delta V/\Delta p$, then below resonance it can be expressed as [9]

$$\delta_m = \frac{s s_0 \sigma_1}{\varepsilon_0(s + \varepsilon s_1)\gamma p_0}. \tag{12.5}$$

It is seen that the sensitivity does not depend on area. If the mass of the membrane is M, then the resonant frequency is defined by

$$fr = \frac{1}{2\pi}\sqrt{\frac{p_0}{s_0 M}}. \tag{12.6}$$

This frequency should be selected well above the upper frequency of the microphone's operating range.

The electret microphone differs from other similar detectors in the sense that it does not require a dc bias voltage. For comparable design dimensions and sensitivity, a condenser microphone would require well over 100 V bias. The mechanical tension

of the membrane is generally kept at a relatively low value (about 10 N m^{-1}), so that the restoring force is determined by the air-gap compressibility. A membrane may be fabricated of Teflon FEP (Fluorinated Ethylene Propylene), which is permanently charged by an electron beam to give it electret properties. The temperature coefficient of sensitivity of the electret microphones are in the range of 0.03 dB/°C in the temperature range from −10 to +50°C [11].

Foil-electret (diaphragm) microphones have more desirable features than any other microphone type. Among them is very wide frequency range from 10^{-3}Hz and up to hundreds of megahertz. They also feature a flat frequency response (within ±1 dB), low harmonic distortion, low vibration sensitivity, good impulse response, and insensitivity to magnetic fields. Sensitivities of electret microphones are in the range of few millivolts per microbar.

For operation in the infrasonic range, an electret microphone requires a miniature pressure equalization hole on the backplate. When used in the ultrasonic range, the electret is often given an additional bias (like a condenser microphone) in addition to its own polarization.

Electret microphones are high-impedance sensors and thus require high-input-impedance interface electronics. A JFET transistor has been the input of choice for many years. However, recently monolithic amplifiers gained popularity. An example is the LMV1014 (National Semiconductors), which is an audio amplifier with very low current consumption (38 μA) that may operate from a small battery power supply ranging from 1.7 to 5 V.

12.6 Solid-State Acoustic Detectors

Currently, use of the acoustic sensors is broader than detecting sound. In particular, they became increasingly popular for detecting mechanical vibrations in a solid for the fabrication of such sensors as microbalances and surface acoustic-wave (SAW) devices. Applications range over measuring displacement, concentration of compounds, stress, force, temperature, and so forth. All such sensors are based on elastic motions in solid parts of the sensor and their major use is serving as parts in other, more complex sensors, (e.g., in chemical detectors, accelerometers, pressure sensors, etc.). In chemical and biological sensors, the acoustic path, where mechanical waves propagate, may be coated with chemically selective compound which interact only with the stimulus of interest.

An excitation device (usually of a piezoelectric nature) forces atoms of the solid into vibratory motions about their equilibrium position. The neighboring atoms then produce a restoring force tending to bring the displaced atoms back to their original positions. In the acoustic sensors, vibratory characteristics, such as phase velocity and/or the attenuation coefficient, are affected by the stimulus. Thus, in acoustic sensors, external stimuli, such as mechanical strain in the sensor's solid, increase the propagating speed of sound. In other sensors, which are called gravimetric, sorption of molecules or attachment of bacteria cause a reduction of acoustic-wave velocity.

In another detector, called the acoustic viscosity sensors, viscous liquid contacts the active region of an elastic-wave sensor and the wave is attenuated.

Acoustic waves propagating in solids have been used quite extensively in electronic devices such as electric filters, delay lines, microactuators, and so forth. The major advantage of the acoustic waves as compared with electromagnetic waves is their low velocity. Typical velocities in solids range from 1.5×10^3 to 12×10^3 m/s, and the practical SAWs utilize the range between 3.8×10^3 and 4.2×10^3 m/s [12], that is, acoustic velocities are five orders of magnitude smaller than those of electromagnetic waves. This allows for the fabrication of miniature sensors operating with frequencies up to 5 GHz.

When the solid-state acoustic sensor is fabricated, it is essential to couple the electronic circuit to its mechanical structure where the waves propagate. The most convenient effect to employ is the piezoelectric effect. The effect is reversible (Section 3.6 of Chapter 3), which means that it works in both directions: The mechanical stress induces electrical polarization charge, and the applied electric field stresses the piezoelectric crystal. Thus, the sensor generally has two piezoelectric transducers at each end: one at the transmitting end, for the generation of acoustic waves, and the other at the receiving end, for conversion of acoustic waves into an electrical signal.

Because silicon does not possess piezoelectric effect, additional piezoelectric material must be deposited on the silicon waver in the form of a thin film [12]. Typical piezoelectric materials used for this purpose are zinc oxide (ZnO), aluminum nitride (AlN), and the so-called solid-solution system of lead–zirconite–titanium oxides $Pb(Zr,Ti)O_3$ known as PZT ceramic. When depositing thin films on the semiconductor material, several major properties must be taken into account:

1. Quality of the adhesion to the substrate
2. Resistance to the external factors (such as fluids which interact with the sensing surface during its operations)
3. Environmental stability (humidity, temperature, mechanical shock, and vibration)
4. Value of electromechanical coupling with the substrate
5. Ease of processing by the available technologies
6. Cost

The strength of the piezoelectric effect in elastic-wave devices depends on the configuration of the transducting electrodes. Depending on the sensor design, for the bulk excitation (when the waves must propagate through the cross-sectional thickness of the sensor), the electrodes are positioned on the opposite sides and their area is quite large. For the SAW, the excitation electrodes are interdigitized.

Several configurations for the solid-state acoustic sensors are known. They differ by the mode the waves propagate through the material. Figure 12.7 shows two of the most common versions: a sensor with a flextural plate mode (Fig. 12.7A) and with the acoustic plate mode (Fig. 12.7B). In the former case, a very thin membrane is flexed by the left pair of the interdigitized electrodes and its vertical deflection induces response in the right pair of the electrodes. As a rule, the membrane thickness is substantially less than the wavelength of the oscillation. In the latter case, the waves are formed on

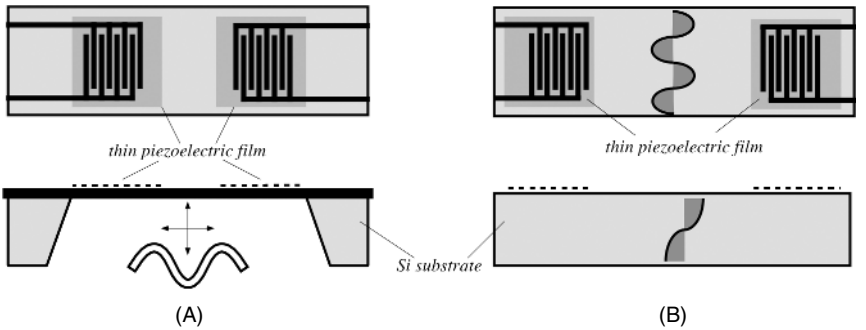


Fig. 12.7. Flextural-plate mode sensor (A) and surface acoustic plate mode (B) sensors.

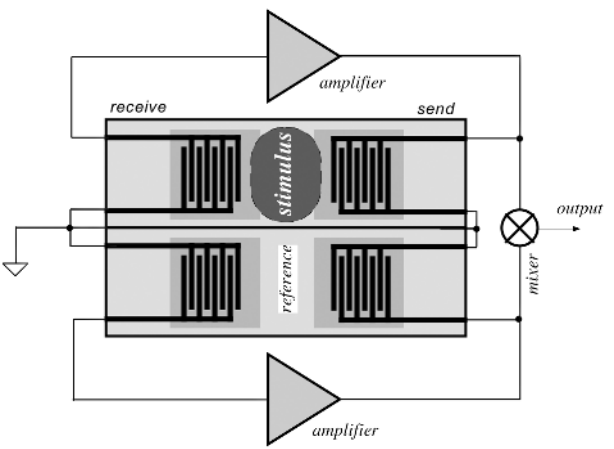


Fig. 12.8. Differential SAW sensor.

the surface of a relatively thick plate. In either case, the space between the left and right pairs of the electrodes is used for interaction with the external stimulus, such as pressure, viscous fluid, gaseous molecules, or microscopic particles.

A typical application circuit for a SAW includes a SAW plate as a time-keeping device of a frequency oscillator. Because many internal and external factors may contribute to the propagation of an acoustic wave and, subsequently, to change in frequency of oscillation, the determination of change in stimulus may be ambiguous and contain errors. An obvious solution is to use a differential technique, where two identical SAW devices are employed: One device is for sensing the stimulus and the other is reference (Fig. 12.8). The reference device is shielded from stimulus, but subjected to common factors, such as temperature, aging, and so forth. The difference of the frequency changes of both oscillators is sensitive only to variations in the stimulus, thus canceling the effects of spurious factors.

References

1. Hohm, D. and Hess, G. A subminiature condenser microphone with silicon nitrite membrane and silicon back plate. *J. Acoust. Soc. Am.* 85, 476–480, 1989.
2. Bergqvist, J. and Rudolf, F. A new condenser microphone in silicon. *Sensors Actuators*, A21–A23, 123–125, 1990.
3. Sprenkels, A.J., Groothengel, R.A., Verloop and A.J., Bergveld, P. Development of an electret microphone in silicon. *Sensor Actuators*, 17(3&4), 509–512, 1989.
4. van der Donk, A.G.H., Sprenkels, A.J., Olthuis, W., and Bergveld, P. Preliminary results of a silicon condenser microphone with internal feedback. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*, IEEE, New York, 1991, pp. 262–265.
5. Wong S.K. and Embleton T.F.W., eds. *AIP Handbook of Condenser Microphones*. AIP Press, New York, 1995.
6. Hellbaum, R.F. et al. An experimental fiber optic microphone for measurement of acoustic pressure levels in hostile environments. In: *Sensors Expo Proceedings*, Helmers Publishing, Peterborough, NH, 1991.
7. *Piezo Film Sensors Technical Manual*. Measurement Specialties, Inc., Norristown, PA, 1999; availabel at www.msiusa.com.
8. Nishikawa, S. and Nukijama, S. *Proc. Imp. Acad. Tokyo* 4, 290, 1928.
9. Sessler, G.M., ed. *Electrets*. Springer-Verlag. Berlin, 1980.
10. Morse, P.M. *Vibration and Sound*. McGraw-Hill, New York, 1948.
11. Griese, H.J., Proc. 9th International Conference on Acoustics, 1977, paper Q29.
12. Motamedi, M.E. and White, R.M. Acoustic sensors. In: *Semiconductor Sensors*. S. M. Sze, ed. John Wiley & Sons, New York, 1994, pp. 97–151.

This page intentionally left blank

13

Humidity and Moisture Sensors

13.1 Concept of Humidity

The water content in surrounding air is an important factor for the well-being of humans and animals. The level of comfort is determined by a combination of two factors: relative humidity and ambient temperature. You may be quite comfortable at −30°C (−22°F) in Siberia, where the air is usually very dry in winter, and feel quite miserable in Cleveland near lake Erie at 0°C (+32°F), where air may contain a substantial amount of moisture.[1] Humidity is an important factor for operating certain equipment (e.g., high-impedance electronic circuits, electrostatic-sensitive components, high-voltage devices, fine mechanisms, etc.). A rule of thumb is to assure a relative humidity near 50% at normal room temperature (20–25°C). This may vary from as low as 38% for the Class-10 clean rooms to 60% in hospital operating rooms. Moisture is the ingredient common to most manufactured goods and processed materials. It can be said that a significant portion of the U.S. GNP (Gross National Product) is moisture [1].

Humidity can be measured by instruments called *hygrometers*. The first hygrometer was invented by Sir John Leslie (1766–1832) [2]. To detect moisture contents, a sensor in a hygrometer must be selective to water, and its internal properties should be modulated by the water concentration. Generally, sensors for moisture, humidity, and dew temperature can be capacitive, conductive, oscillating, or optical. The optical sensors for gases detect dew-point temperature, whereas the optical hygrometers for organic solvents employ absorptivity of near-infrared (NIR) light in the spectral range from 1.9 to 2.7 μm [3] (see Fig. 14.18 of Chapter 14).

There are many ways to express moisture and humidity, often depending on the industry or the particular application. The moisture of gases is expressed sometimes in pounds of water vapor per million cubic feet of gas. The moisture in liquids and solids is generally given as a percentage of water per total mass (wet-weight basis), but may be given on a dry-weight basis. The moisture in liquids with low water miscibility is usually expressed as parts per million by weight (PPM_w).

[1] Naturally, here we disregard other comfort factors, such as economical, cultural, and political.

The term *moisture* generally refers to the water content of any material, but for practical reasons, it is applied only to liquids and solids, whereas the term *humidity* is reserved for the water vapor content in gases. The following are some useful definitions:

Moisture: the amount of water contained in a liquid or solid by absorption or adsorption which can be removed without altering its chemical properties.

Mixing ratio (humidity ratio) r: the mass of water vapor per unit mass of dry gas.

Absolute humidity (mass concentration or density of water vapor): the mass m of water vapor per unit volume v of wet gas: $d_w = m/v$. In other words, absolute humidity is the density of the water vapor component. It can be measured, for example, by passing a measured quantity of air through a moisture-absorbing substance (such as silica gel) which is weighed before and after the absorption. Absolute humidity is expressed in grams per cubic meter, or in grains per cubic foot. Because this measure is also a function of atmospheric pressure, it is not generally useful in engineering practice.

Relative humidity: the ratio of the actual vapor pressure of the air at any temperature to the maximum of saturation vapor pressure at the same temperature. Relative humidity in percent is defined as

$$H = 100\frac{P_w}{P_s}, \tag{13.1}$$

where P_w is the partial pressure of water vapor and P_s is the pressure of saturated water vapor at a given temperature. The value of H expresses the vapor content as a percentage of the concentration required to cause the vapor saturation, [i.e., the formation of water droplets (dew) at that temperature]. An alternative way to present RH is as a ratio of the mole fraction of water vapor in a space to the mole fraction of water vapor in the space at saturation. The value of P_w together with partial pressure of dry air P_a is equal to pressure in the enclosure, or to the atmospheric pressure P_{atm} if the enclosure is open to the atmosphere:

$$P_w + P_a = P_{\text{atm}}. \tag{13.2}$$

At temperatures above the boiling point, water pressure could displace all other gases in the enclosure. The atmosphere would then consist entirely of superheated steam. In this case, $P_w = P_{\text{atm}}$. At temperatures above 100°C, RH is a misleading indicator of moisture content because at these temperatures P_s is always more than P_{atm}, and maximum RH can never reach 100%. Thus, at normal atmospheric pressure and a temperature of 100°C, the maximum RH is 100%, whereas at 200°C, it is only 6%. Above 374°C, saturation pressures are not thermodynamically specified.

Dew-point temperature: the temperature at which the partial pressure of the water vapor present would be at its maximum, or saturated vapor condition, with respect to equilibrium with a plain surface of ice. It also is defined as the temperature to which the gas–water vapor mixture must be cooled isobarically (at constant

Table 13.1. Relative Humidity of Saturated Salt Solutions

Temperature (°C)	Lithium Chloride Solution (LiCl, H_2O)	Magnesium Chloride Solution (MgCl, $6H_2O$)	Magnesium Nitrate Solution ($Mg(NO_3)_2$, $6H_20$)	Sodium Chloride Solution (NaCl, $6H_2O$)	Potassium Chloride Solution K_2SO_4
5	13	33.6 ± 0.3	58	75.7 ± 0.3	98.5 ± 0.9
10	13	33.5 ± 0.2	57	75.7 ± 0.2	98.2 ± 0.8
15	12	33.3 ± 0.2	56	75.6 ± 0.2	97.9 ± 0.6
20	12	33.1 ± 0.2	55	75.5 ± 0.1	97.6 ± 0.5
25	**11.3 ± 0.3**	**32.8 ± 0.3**	**53**	**75.3 ± 0.1**	**97.3 ± 0.5**
30	11.3 ± 0.2	32.4 ± 0.1	52	75.1 ± 0.1	97.0 ± 0.4
35	11.3 ± 0.2	32.1 ± 0.1	50	74.9 ± 0.1	96.7 ± 0.4
40	11.2 ± 0.2	31.6 ± 0.1	49	74.7 ± 0.1	96.4 ± 0.4
45	11.2 ± 0.2	31.1 ± 0.1	—	74.5 ± 0.2	96.1 ± 0.4
50	11.1 ± 0.2	30.5 ± 0.1	46	74.6 ± 0.9	95.8 ± 0.5
55	11.0 ± 0.2	29.9 ± 0.2	—	74.5 ± 0.9	—

Source: Patissier, B., Walters, D. Basics of relative humidity calibration for Humirel HS1100/HS1101 sensors. Humirel, Toulouse Cedex, 1999.

pressures) to induce frost or ice (assuming no prior condensation). The *dew point* is the temperature at which relative humidity is 100%. In other words, the dew point is the temperature that the air must reach for the air to hold the maximum amount of moisture it can. When the temperature cools to the dew point, the air becomes saturated and fog, dew, or frost can occur.

The following equations [4] calculate the dew point from relative humidity and temperature. All temperatures are in Celsius. The saturation vapor pressure over water is found from

$$\mathrm{EW} = 10^{0.66077+7.5t/(237+t)} \tag{13.3}$$

and the dew-point temperature is found from the approximation

$$\mathrm{DP} = \frac{237.3\left(0.66077 - \log_{10}\mathrm{EW_{RH}}\right)}{\log_{10}\mathrm{EW_{RH}} - 8.16077}t \tag{13.4}$$

where

$$\mathrm{EW_{RH}} = \frac{(\mathrm{EW}) \cdot (\mathrm{RH})}{100}.$$

The relative humidity displays an inverse relationship with the absolute temperature. The dew-point temperature is usually measured with a chilled mirror. However, below the 0°C dew point, the measurement becomes uncertain, as moisture eventually freezes and a crystal lattice growth will slowly occur, much like a snowflake. Nevertheless, moisture can exist for a prolonged time below 0°C in a liquid phase, depending on such variables as molecular agitation, rate of convection, sample gas temperature, contaminations, and so forth.

13.2 Capacitive Sensors

An air-filled capacitor may serve as a relative humidity sensor because moisture in the atmosphere changes air electrical permitivity according to the following equation [5]:

$$\kappa = 1 + \frac{211}{T}\left(P + \frac{48P_s}{T}H\right)10^{-6} \tag{13.5}$$

where T is the absolute temperature (in K), P is the pressure of moist air (in mm Hg), P_s is the pressure of saturated water vapor at temperature T (in mm Hg), H is the relative humidity (in %). Equation (13.5) shows that the dielectric constant of moist air and, therefore, the capacitance are proportional to the relative humidity.

Instead of air, the space between the capacitor plates can be filled with an appropriate isolator whose dielectric constant changes significantly upon being subjected to humidity. The capacitive sensor may be formed of a hygroscopic polymer film with metallized electrodes deposited on the opposite sides. In one design [6], the dielectric was composed of a hygrophilic polymer thin film (8–12 μm thick) made of cellulose acetate butyrate and the dimetylephtalate as plasticizer. The size of the film sensor is 12 × 12 mm. The 8-mm-diameter gold porous disk electrodes (200 Å thick) were deposited on the polymer by vacuum deposition. The film was suspended by a holder and the electrodes were connected to the terminals. The capacitance of such a sensor is approximately proportional to relative humidity H

$$C_h \approx C_0(1 + \alpha_h H), \tag{13.6}$$

where C_0 is the capacitance at $H = 0$.

For the use with capacitive sensors, a 2% accuracy in the range from 5% to 90% RH can be achieved with a simple circuit as shown in Fig. 13.1. The sensor and the circuit transfer characteristics are shown in Fig. 13.2. The sensor's nominal capacitance at 75% RH is 500 pF. It has a quasilinear transfer function with the offset at zero humidity of about 370 pF and a slope of 1.7 pF/% RH. The circuit effectively performs two functions: makes a capacitance-to-voltage conversion and subtracts the offset capacitance to produce an output voltage with zero intercept. The heart of the circuit is a self-clocking analog switch LT1043, which multiplexes several capacitors at the summing junction (virtual ground) of the operational amplifier U_1. The capacitor C_1 is for the offset capacitance subtraction, whereas the capacitor C_2 is connected in series with the capacitive sensor S_1. The average voltage across the sensor must be zero; otherwise, electrochemical migration could damage it permanently. The nonpolarized capacitor C_2 protects the sensor against building up any dc charge. Trimpot P_2 adjusts the amount of charge delivered to the sensor and P_1 trims the offset charge which is subtracted from the sensor. The net charge is integrated with the help of the feedback capacitor C_3. Capacitor C_4 maintains the dc output when the summing junction is disconnected from the sensor.

A similar technique can be used for measuring moisture in material samples [7]. Figure 13.3 shows a block diagram of the capacitive measurement system where the dielectric constant of the sample changes the frequency of the oscillator. This method

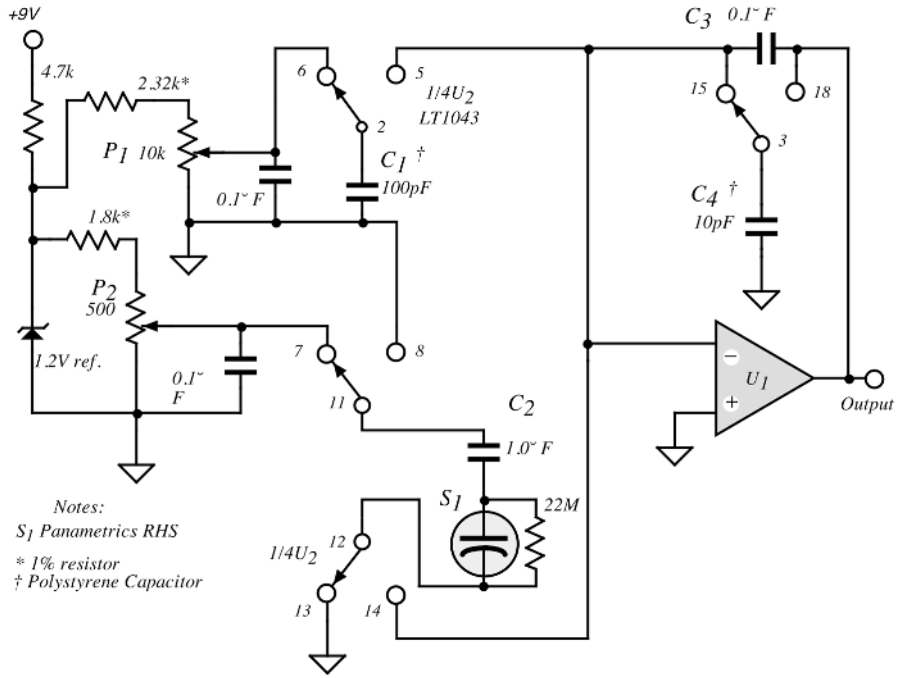


Fig. 13.1. Simplified circuit for measuring humidity with a capacitive sensor. (Adapted from Ref. [6].)

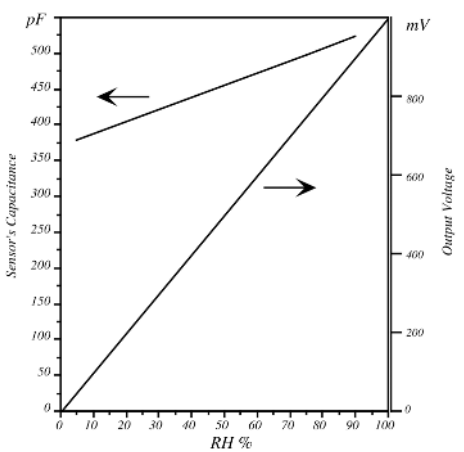


Fig. 13.2. Transfer functions of a capacitive sensor and a system.

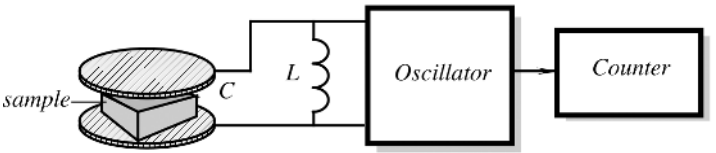


Fig. 13.3. Capacitive moisture sensing system.

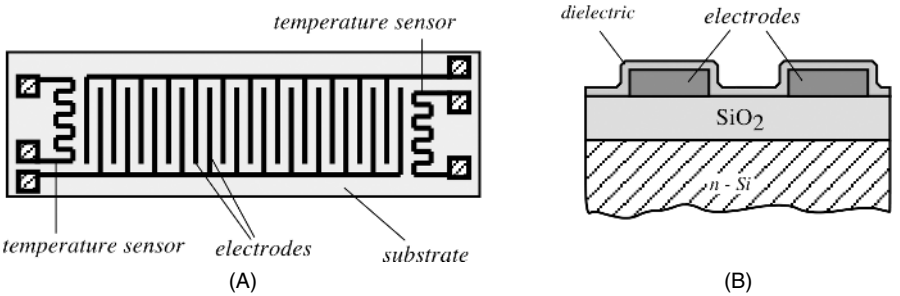


Fig. 13.4. Capacitive thin-film humidity sensor: (A) interdigitized electrodes form capacitor plates; (B) cross section of the sensor.

of moisture measurement is quite useful in the process control of pharmaceutical products. The dielectric constants of most medical tablets is quite low (between 2.0 and 5.0) as compared with that of water (Fig. 3.7 of Chapter 3). The sampled material is placed between two test plates which form a capacitor connected into an LC oscillating circuit. The frequency is measured and related to the moisture. The best way to reduce variations attributed to environmental conditions, such as temperature and room humidity, is the use of a differential technique; that is, the frequency shift $\Delta f = f_0 - f_1$ is calculated, where f_0 and f_1 are frequencies produced by the empty container and that filled with the sampled material, respectively. The method has some limitations; for instance, its accuracy is poor when measuring moistures below 0.5%, the sample must be clean of foreign particles having relatively high dielectric constants (e.g., metal and plastic objects, a packing density), and a fixed sample geometry must be maintained.

A thin-film capacitive humidity sensor can be fabricated on a silicon substrate [8]. A layer of SiO_2 3000 Å thick is grown on an n-Si substrate (Fig. 13.4B). Two metal electrodes are deposited on the SiO_2 layer. They are made of aluminum, chromium, or phosphorus-doped polysilicon (LPCVD).[2] The electrode thickness is in the range 2000–5000 Å. The electrodes are shaped in an interdigitized pattern as shown in Fig. 13.4A. To provide additional temperature compensation, two temperature-sensitive resistors are formed on the same substrate. The top of the sensor is coated with a dielectric layer. For this layer, several materials can be used, such as chemically

[2] LPCVD - low pressure chemical vapor deposition.

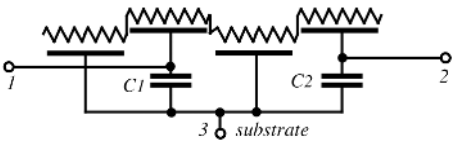


Fig. 13.5. Simplified equivalent electric circuit of a capacitive thin-film humidity sensor.

vapor-deposited SiO_2 or phosphorosilicate glass (CVD PSG). The thickness of the layer is in the range 300–4000 Å.

A simplified equivalent electrical circuit is shown in Fig. 13.5. Each element of the circuit represents a *RC* transmission line [9]. When the relative humidity increases, the distributed surface resistance drops and the equivalent capacitance between terminals 1 and 2 grows. The capacitance is frequency dependent; hence, for the low-humidity-range measurement, the frequency should be selected near 100 Hz, whereas for the higher humidities, it is in the range between 1 and 10 kHz.

13.3 Electrical Conductivity Sensors

Resistances of many nonmetal conductors generally depend on their water content, as was discussed in Section 3.5.4 of Chapter 3. This phenomenon is the basis of a resistive humidity sensor or hygristor. A general concept of a conductive hygrometric sensor is shown in Fig. 13.6. The sensor contains a material of relatively low resistivity which changes significantly under varying humidity conditions. The material is deposited on the top of two interdigitized electrodes to provide a large contact area. When water molecules are absorbed by the upper layer, resistivity between the electrodes changes and can be measured by an electronic circuit. The first such sensor was developed by F. W. Dunmore in 1935; it was a hygroscopic film consisting of 2–5% aqueous solution of LiCl [10]. Another example of a conductive humidity sensor is the so-called "Pope element," which contains a polystyrene film treated with sulfuric acid to obtain the desired surface-resistivity characteristics.

Other promising materials for the fabrication of a film in a conductivity sensor are solid polyelectrolytes because their electrical conductivity varies with humidity. Long-term stability and repeatability of these compounds, although generally not too

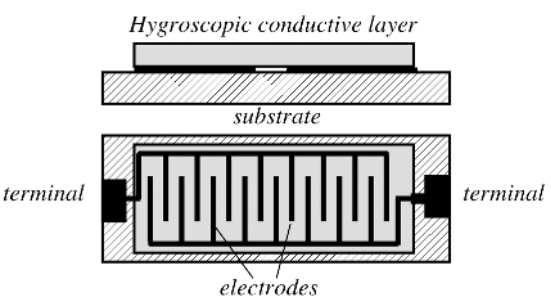


Fig. 13.6. Composition of a conductive humidity sensor.

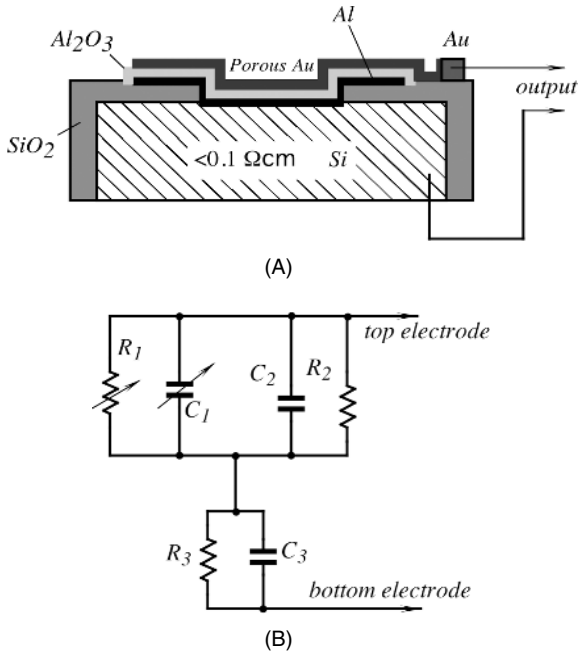


Fig. 13.7. (A) Structure of Al_2O_3 thin film moisture sensor; (B) simplified equivalent circuit of the sensor. R_1 and C_1 are moisture-dependent variable terms; R_2 and C_2 are shunting terms of bulk oxide between pores (unaffected by moisture); R_3 and C_3 are series terms below pores (unaffected by moisture).

great, can be significantly improved by using the interpenetrating polymer networks and carriers and supporting media. When measured at 1 kHz, an experimental sample of such a film has demonstrated a change in impedance from 10 MΩ to 100 Ω as the RH changed from 0% to 90% [11].

A solid-state humidity sensor can be fabricated on a silicon substrate (Fig. 13.7A). The silicon must be of a high conductance [12], which provides an electrical path from the aluminum electrode vacuum deposited on its surface. An oxide layer is formed on the top of the conductive aluminum layer, and on the top of that, another electrode is formed. The aluminum layer is anodized in a manner to form a porous oxide surface. The average cross-sectional dimension of pores is sufficient to allow penetration by water molecules. The upper electrode is made of a form of porous gold which is permeable to gas and, at the same time, can provide electrical contact. Electrical connections are made to the gold and silicon layers. Aluminum oxide (Al_2O_3), like numerous other materials, readily sorb water when in contact with a gas mixture containing water in the vapor phase. The amount of sorption is proportional to the water vapor partial pressure and inversely proportional to the absolute temperature. Aluminum oxide is a dielectric material. Its dielectric constant and surface resistivity are modified by the physisorption of water. For this reason, this material can be used as a humidity sensing compound.

Figure 13.7B shows an electrical equivalent circuit of the sensor [13]. The values of R_1 and C_1 depend on the Al_20_3 average pore sizes and density. These components of resistance and capacitance vary with the number of water molecules that penetrate the pores and adhere to the surface. R_2 and C_2 represent the resistance and capacitance components of the bulk oxide material between the pores and are therefore unaffected by moisture. C_3 is an equivalent series capacitance term as determined by the measurement of the total resistance components in a dry atmosphere at very low frequencies. The sensor's resistance becomes very large ($> 10^8\Omega$) as the frequency approaches dc. Thus, the measurement of humidity involves the measurement of the sensor's impedance. The residual of nonhumidity-dependent resistance and capacitance terms that exist in a typical sensor shunt the humidity-dependent variables, thus causing the continuous reduction in slope (sensitivity) as the humidity is lowered, which, in turn, reduces the accuracy at lower humidities. Because temperature is a factor in humidity measurement, the sensor usually combines a humidity sensor, a thermistor, and a reference capacitance in the same package, which is protected against humidity influence and has a low-temperature coefficient.

13.4 Thermal Conductivity Sensor

Using the thermal conductivity of gas to measure humidity can be accomplished by a thermistor-based sensor (Fig. 13.8A) [14]. Two tiny thermistors (R_{t1} and R_{t2}) are supported by thin wires to minimize thermal conductivity loss to the housing. The left thermistor is exposed to the outside gas through small venting holes, and the right thermistor is hermetically sealed in dry air. Both thermistors are connected into a bridge circuit (R_1 and R_2), which is powered by voltage $+E$. The thermistors develop self-heating due to the passage of electric current. Their temperatures rise up to 170°C over the ambient temperature. Initially, the bridge is balanced in dry air to establish a zero reference point. The output of this sensor gradually increases as

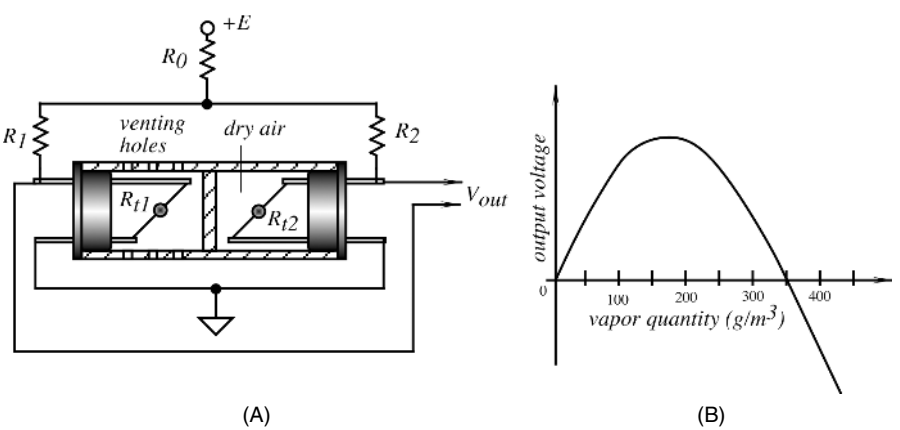


Fig. 13.8. Absolute humidity sensor with self-heating thermistors: (A) design and electrical connection; (B) output voltage.

absolute humidity rises from zero. At about 150 g/m^3, it reaches the saturation and then decreases with a polarity change at about 345 g/m^3 (Fig. 13.8B).

13.5 Optical Hygrometer

Most of the humidity sensors exhibit some repeatability problems, especially hysteresis with a typical value from 0.5% to 1% RH. In precision process control, this may be a limiting factor; therefore indirect methods of humidity measurements should be considered. The most efficient method is a calculation of absolute or relative humidity through the dew-point temperature. As indicated earlier, the *dew point* is the temperature at which liquid and vapor phases of water (or any fluid for that matter) are in equilibrium. The temperature at which the vapor and solid phases are in equilibrium is called the *frost point*. At the dew point, only one value of saturation vapor pressure exists. Hence, absolute humidity can be measured from this temperature as long as the pressure is known. The optimum method of moisture measurement by which the minimum hysteresis effects are realized requires the use of optical hygrometry. The cost of an optical hygrometer is considerably greater, but if the benefit of tracking low-level moisture enhances product yield and quality, the cost is easily justified.

The basic idea behind the optical hygrometer is the use of a mirror whose surface temperature is precisely regulated by a thermoelectric heat pump. The mirror temperature is controlled at the threshold of the formation of dew. Sampled air is pumped over the mirror surface, and if the mirror temperature crosses a dew point, it releases moisture in the form of water droplets. The reflective properties of the mirror change at water condensation because water droplets scatter light rays. This can be detected by an appropriate photodetector. Figure 13.9 shows a simplified block diagram of a

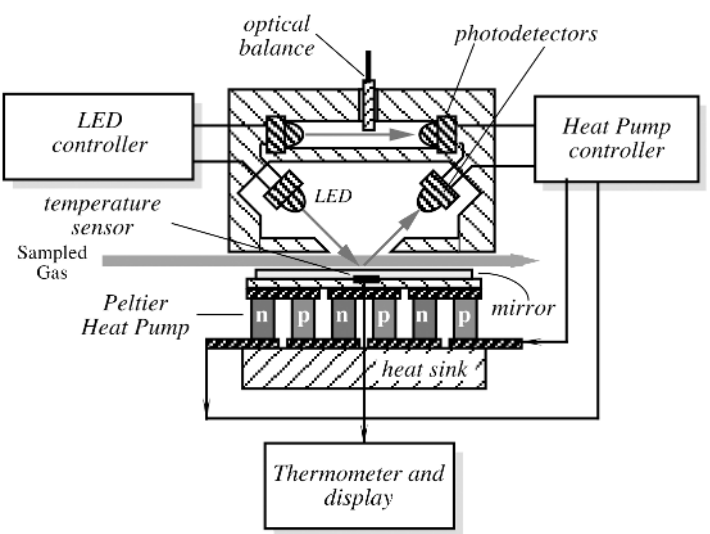


Fig. 13.9. Chilled-mirror dew-point sensor with an optical bridge.

chilled-mirror hygrometer. It is comprised of a heat pump operating on the Peltier effect. The pump removes heat from a thin mirrored surface which has an imbedded temperature sensor. That sensor is part of a digital thermometer which displays the temperature of the mirror. The hygrometer's circuit is of a differential type, where the top optocoupler, a light-emitting diode (LED) and a photodetector, are used for the compensation of drifts; the bottom optocoupler is for measuring the mirror's reflectivity. The sensor's symmetry can be balanced by a wedged optical balance inserted into the light path of the upper optocoupler. The lower optocoupler is positioned at a 45° angle with respect to the mirror. Above the dew point, the mirror is dry and its reflectivity is the highest. The heat pump controller lowers the temperature of the mirror through the heat pump. At the moment of water condensation, the mirror reflectivity drops abruptly, which causes the reduction in the photocurrent in the photodetector. The photodetector's signals pass to the controller to regulate electric current through the heat pump to maintain its surface temperature at the level of the dew point, where no additional condensation or evaporation from the mirror surface occurs. Actually, water molecules are continuously being trapped and are escaping from the surface, but the average net level of the condensate density does not change once equilibrium is established.

Because the sensed temperature of the mirrored surface precisely determines the actual prevailing dew point, this is considered the moisture's most fundamental and accurate method of measurement. Hysteresis is virtually eliminated and sensitivity is near 0.03°C DP (dew point). From the dew point, all moisture parameters such as %RH, vapor pressure, and so forth. are obtainable once the prevailing temperature and pressure are known.

There are several problems associated with the method. One is a relatively high cost, the other is potential mirror contamination, and the third is a relatively high power consumption by the heat pump. Contamination problems can be virtually eliminated with use of particle filters and a special technique that deliberately cools the mirror well below the dew point to cause excessive condensation with the following fast rewarming. This flushes the contaminants, keeping the mirror clean [15].

13.6 Oscillating Hygrometer

The idea behind the oscillating hygrometer is similar to that behind the optical chilled-mirror sensor. The difference is that the measurement of the dew point is made not by the optical reflectivity of the surface, but rather by detecting the changing mass of the chilled plate. The chilled plate is fabricated of a thin quartz crystal that is a part of an oscillating circuit. This implies the other name for the sensor: *the piezoelectric hygrometer*, because the quartz plate oscillation is based on the piezoelectric effect. A quartz crystal is thermally coupled to the Peltier cooler (see Section 3.9 of Chapter 3), which controls the temperature of the crystal with a high degree of accuracy (Fig. 13.10). When the temperature drops to that of a dew point, a film of water vapor deposits on the exposed surface of the quartz crystal. Because the mass of the crystal changes, the resonant frequency of the oscillator shifts from f_0 to f_1. The new frequency f_1 corresponds to a given thickness of the water layer. The frequency

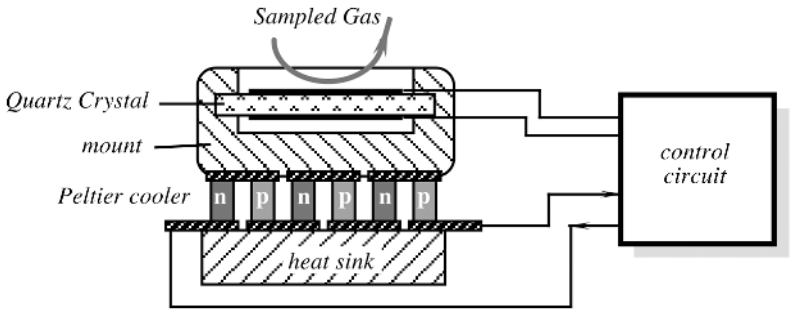


Fig. 13.10. Oscillating humidity sensor.

shift controls current through the Peltier cooler, thus changing the temperature of the quartz crystal to stabilize at the dew point temperature. The major difficulty in designing the piezoelectric hygrometer is in providing an adequate thermal coupling between the cooler and the crystal while maintaining small size of the crystal at a minimum mechanical loading [16]. Naturally, this method may be employed by using the surface acoustic-wave (SAW) sensors, similar to that of Fig. 12.8 of Chapter 12, where the stimulus place is the area subjected to the sampled gas.

References

1. Quinn, F. C. The most common problem of moisture/humidity measurement and control. In: *Moisture and Humidity, Proceedings of the 1985 International Symposium on Moisture and Humidity*. Chaddock, J. B. ed., ISA, Washington, DC, 1985, pp. 1–5.
2. Carter, E. F., ed. *Dictionary of Inventions and Discoveries*. Crane, Russak and Co., New York, 1966.
3. Baughman E. H. and Mayes, D. NIR applications to process analysis. *Am. Lab.*, 21(10), 54–58, 1989.
4. Berry, F. A., Jr. *Handbook of Meteorology*. McGraw-Hill, New York, 1945, p. 343.
5. *Conditioner Circuit*, Appl. Handbook, Linear Technology, Inc., Milpitas, 1990.
6. Sashida, T. and Sakaino, Y. An interchangeable humidity sensor for an industrial hygrometer. In: *Moisture and Humidity. Proceedings of the International Symposium on Moisture and Humidity*. Chaddock, J. B. ed., Washington, DC, 1985.
7. Carr-Brion, K. *Moisture Sensors in Process Control*. Elsevier Applied Science, New York, 1986.
8. Jachowicz, R. S. and Dumania, P. Evaluation of thin-film humidity sensor type MCP–MOS. In: *Moisture and Humidity. Proceedings of the International Symposium on Moisture and Humidity*. Chaddock, J. B. ed., ISA, Washington, DC, 1985.

9. Jachowicz, R. S. and Senturia, S.D. A thin film humidity sensor. *Sensors Actuators* 2, 1981, 1982.
10. Norton, H. N. *Handbook of Transducers*. Prentice-Hall, Englewood Cliffs, NJ, 1989.
11. Sakai, Y., Sadaoka, Y., Matsuguchi, M., and Hirayama, K. Water resistive humidity sensor composed of interpenetrating polymer networks of hydrophilic and hydrophobic methacrylate. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*, IEEE, New York, 1991, pp. 562–565.
12. Fong, V. Al_2O_3 moisture sensor chip for inclusion in microcircuit package and the new MIL standard for moisture content. In: *Moisture and Humidity, Proceedings of the 1985 International Symposium on Moisture and Humidity*. Chaddock, J. B., ed., ISA, Washington, DC, 1985, pp. 345–357.
13. Harding, J. C., Jr., Overcoming limitations inherent to aluminum oxide humidity sensors. In: *Moisture and Humidity, Proceedings of the 1985 Internation Symposium on Moisture and Humidity*. Chaddock, J. B., ed., ISA, Washington, DC, 1985, pp. 367–378.
14. Miura, T. Thermistor humidity sensor for absolute humidity measurements and their applications. In: *Moisture and Humidity. Proceedings of the International Symposium on Moisture and Humidity*, Chaddock, J. B., ed., ISA, Washington, DC, 1985.
15. Harding, J. C., Jr., A chilled mirror dewpoint sensor/psychrometric transmitter for energy monitoring and control systems. In: *Moisture and Humidity. Proceeding of the International Symposium on Moisture and Humidity*, Chaddock, J. B., ed., ISA, Washington, DC, 1985.
16. Porlier, C. Chilled piezoelectric hygrometer: sensor interface design. In: *Sensors Expo Proceedings*. Helmers Publishing, 1991, paper 107B-7.

This page intentionally left blank

14

Light Detectors

"There is nothing more practical than a good theory"

—Gustav Robert Kirchhoff

14.1 Introduction

Detectors of electromagnetic radiation in the spectral range from ultraviolet to far infrared are called light detectors. From the standpoint of a sensor designer, absorption of photons by a sensing material may result either in a quantum or thermal response. Therefore, all light detectors are divided into two major groups that are called *quantum* and *thermal*. The quantum detectors operate from the ultraviolet to mid-infrared spectral ranges, whereas thermal detectors are most useful in the mid- and far-infrared spectral range where their efficiency at room temperatures exceeds that of the quantum detectors. In this chapter, we cover both types. For a description of highly sensitive photon sensors called *photomultipliers*, refer to Section 15.1 of Chapter 15.

Solid-state quantum detectors (photovoltaic and photoconductive devices) rely on the interaction of individual photons with a crystalline lattice of semiconductor materials. Their operations are based on the photoeffect that was discovered by A. Einstein, and brought him the Nobel Prize. In 1905, he made a remarkable assumption about the nature of light: that, at least under certain circumstances, its energy was concentrated into localized bundles, later named photons. The energy of a single photon is given by

$$E = hv, \tag{14.1}$$

where v is the frequency of light and $h = 6.626075 \times 10^{-34}$ J s (or 4.13567×10^{-15}eVs) is Planck's constant derived on the basis of the wave theory of light. When a photon strikes the surface of a conductor, it may result in the generation of a free electron. Part (ϕ) of the photon energy E is used to detach the electron from the surface; the other part gives its kinetic energy to the electron. The photoelectric effect can be described as

$$hv = \phi + K_m, \tag{14.2}$$

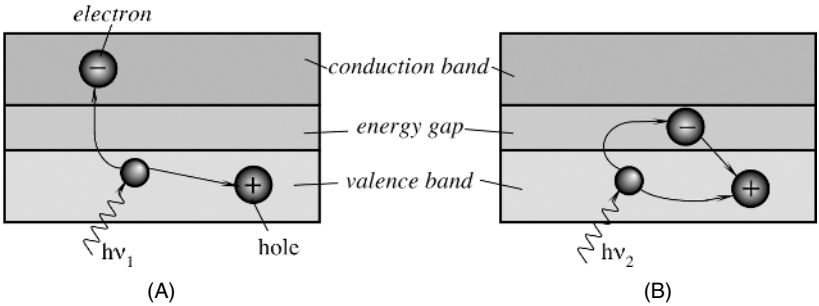


Fig. 14.1. Photoeffect in a semiconductor for high-energy (A) and low-energy (B) photons.

where ϕ is called the *work function* of the emitting surface and K_m is the maximum kinetic energy of the electron upon its exiting the surface. Similar processes occur when a semiconductor p-n junction is subjected to radiant energy: The photon transfers its energy to an electron, and if the energy is sufficiently high, the electron may become mobile, which results in an electric current.

The periodic lattice of crystalline materials establishes allowed energy bands for electrons that exist within that solid. The energy of any electron within the pure material must be confined to one of these energy bands, which may be separated by gaps or ranges of forbidden energies.

If light of a proper wavelength [sufficiently high energy of photons; see Eq. (14.1)] strikes a semiconductor crystal, the concentration of charge carriers (electrons and holes) in the crystal increases, which manifests in the increased conductivity of a crystal:

$$\sigma = e(\mu_e n + \mu_h p), \tag{14.3}$$

where e is the electron charge, μ_e is the electron mobility, μ_h is the hole mobility, and n and p are the respective concentrations of electrons and holes.

Figure 14.1A shows energy bands of a semiconductor material, where E_g is the magnitude in electron volts (eV) of the forbidden band gap. The lower band is called the valence band, which corresponds to those electrons that are bound to specific lattice sites within the crystal. In the case of silicon or germanium, they are parts of the covalent bonding which constitute the interatomic forces within the crystal. The next higher-lying band is called the *conduction band* and represents electrons that are free to migrate through the crystal. Electrons in this band contribute to the electrical conductivity of the material. The two bands are separated by the band gap, the size of which determines whether the material is classified as a semiconductor or an isolator. The number of electrons within the crystal is just adequate to completely fill all available sites within the valence band. In the absence of thermal excitation, both isolators and semiconductors would therefore have a configuration in which the valence band is completely full and the conduction band completely empty. Under these imaginable circumstances, neither would theoretically show any electrical conductivity.

In a metal, the highest occupied energy band is not completely full. Therefore, electrons can easily migrate throughout the material because they need achieve only

Table 14.1. Band Gaps and Longest Wavelengths for Various Semiconductors

Material	Band Gap (eV)	Longest Wavelength (μm)
ZnS	3.6	0.345
CdS	2.41	0.52
CdSe	1.8	0.69
CdTe	1.5	0.83
Si	1.12	1.10
Ge	0.67	1.85
PbS	0.37	3.35
InAs	0.35	3.54
Te	0.33	3.75
PbTe	0.3	4.13
PbSe	0.27	4.58
InSb	0.18	6.90

Source: Ref. [1].

a small incremental energy to the above occupied states. Metals, therefore, are always characterized by a very high electrical conductivity. In isolators or semiconductors, on the other hand, the electron must first cross the energy band gap in order to reach the conduction band and the conductivity is, therefore, many orders of magnitude lower. For isolators, the band gap is usually 5 eV or more, whereas for semiconductors, the gap is considerably less (Table 14.1). Note that the longer the wavelength (lower frequency of a photon), the less energy is required to originate a photoeffect.

When the photon of frequency v_1 strikes the crystal, its energy is high enough to separate the electron from its site in the valence band and push it through the band gap into a conduction band at a higher energy level. In that band, the electron is free to serve as a current carrier. The deficiency of an electron in the valence band creates a hole which also serves as a current carrier. This is manifested in the reduction of specific resistivity of the material. On the other hand, Fig. 14.1B shows that a photon of lower frequency v_2 does not have sufficient energy to push the electron through the band gap. The energy is released without creating current carriers.

The energy gap serves as a threshold below which the material is not light sensitive. However, the threshold is not abrupt. Throughout the photon-excitation process, the law of conservation of momentum applies. The momentum and density of hole–electron sites are higher at the center of both the valence and conduction bands, and they fall to zero at the upper and lower ends of the bands. Therefore, the probability of an excited valence-band electron finding a site of like momentum in the conduction band is greater at the center of the bands and is the lowest at the ends of the bands. Therefore, the response of a material to photon energy increases from E_g gradually to its maximum and then falls back to zero at the energy corresponding to the difference between the bottom of the valence band and the top of the conduction band. A typical spectral response of a semiconductive material is shown in Fig. 14.2. The light response of a bulk material can be altered by adding various impurities. They can be

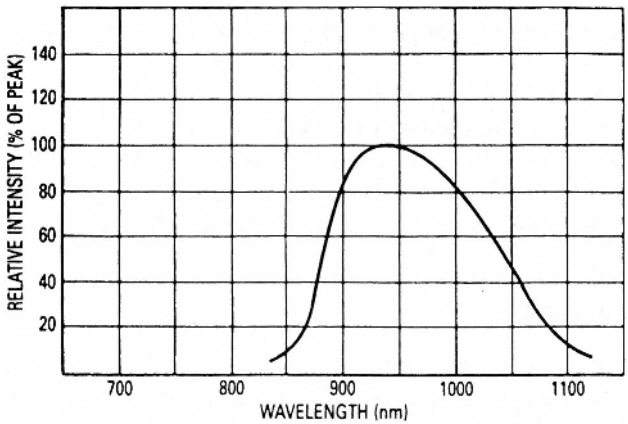


Fig. 14.2. Spectral response of an infrared photodiode.

used to reshape and shift a spectral response of the material. All devices that directly convert photons of electromagnetic radiation into charge carriers are called *quantum detectors* and are generally produced in a form of photodiodes, phototransistors, and photoresistors.

When comparing the characteristics of different photodetectors, the following specifications usually should be considered:

NEP (noise-equivalent power) is the amount of light equivalent to the intrinsic noise level of the detector. Stated differently, it is the light level required to obtain a signal-to-noise ratio equal to unity. Because the noise level is proportional to the square root of the bandwidth, the NEP is expressed in units of W/$\sqrt{\text{Hz}}$:

$$\text{NEP} = \frac{\text{noise current (A}/\sqrt{\text{HZ}})}{\text{Radiant sensitivity at } \lambda_p \text{ (A/W)}}. \tag{14.4}$$

D^* refers to the *detectivity* of a detector's sensitive area of 1 cm^2 and a noise bandwidth of 1 Hz:

$$D^* = \frac{\sqrt{\text{Area(cm}^2)}}{\text{NEP}}. \tag{14.5}$$

Detectivity is another way of measuring the sensor's signal-to-noise ratio. Detectivity is not uniform over the spectral range for operating frequencies; therefore, the chopping frequency and the spectral content must be also specified. The detectivity is expressed in units of cm $\sqrt{\text{Hz}}$/W. It can be said that the higher the value of D^*, the better the detector.

IR cutoff wavelength (λ_c) represents the long-wavelength limit of the spectral response and often is listed as the wavelength at which the detectivity drops by 10% of the peak value.

Maximum current is specified for photoconductive detectors (such as HgCdTe) which operate at constant currents. The operating current never should exceed the maximum limit.

Maximum reverse voltage is specified for Ge and Si photodiodes and photoconductive cells. Exceeding this voltage can cause the breakdown and severe deterioration of the sensor's performance.

Radiant responsivity is the ratio of the output photocurrent (or output voltage) divided by the incident radiant power at a given wavelength, expressed in A/W or V/W.

Field of view (FOV) is the angular measure of the volume of space where the sensor can respond to the source of radiation.

Junction capacitance (C_j) is similar to the capacitance of a parallel-plate capacitor. It should be considered whenever a high-speed response is required. The value of C_j drops with reverse bias and is higher for the larger diode areas.

14.2 Photodiodes

Photodiodes are semiconductive optical sensors, which, if broadly defined, may even include solar batteries. However, here we consider only the information aspect of these devices rather than the power conversion. In a simple way, the operation of a photodiode can be described as follows. If a p-n junction is forward biased (positive side of a battery is connected to the *p* side) and is exposed to light of proper frequency, the current increase will be very small with respect to a dark current. In other words, the bias current is much greater than the current generated by light. If the junction is *reverse biased* (Fig. 14.3), the current will increase quite noticeably. Impinging photons create electron–hole pairs on both sides of the junction. When electrons enter the conduction band, they start flowing toward the positive side of the battery. Correspondingly, the created holes flow to the negative terminal, meaning that the

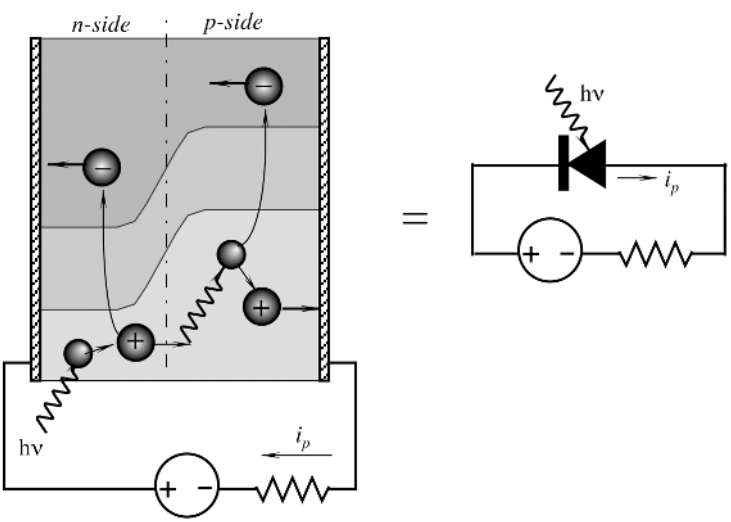


Fig. 14.3. Structure of a photodiode.

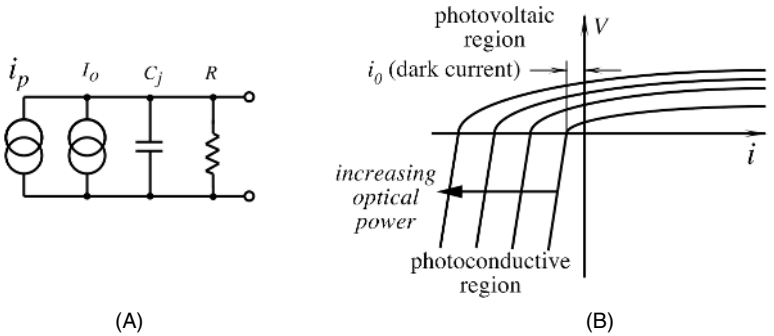


Fig. 14.4. An equivalent circuit of a photodiode (A) and its volt-ampere characteristic (B).

photocurrent i_p flows in the network. Under dark conditions, the leakage current i_0 is independent of applied voltage and mainly is the result of thermal generation of charge carriers. Thus, a reverse-biased photodiode electrical equivalent circuit (Fig. 14.4A) contains two current sources and an RC network.

The process of optical detection involves the direct conversion of optical energy (in the form of photons) into an electrical signal (in the form of electrons). If the probability that a photon of energy hv will produce an electron in a detection is η, then the average rate of production of electrons $\langle r \rangle$ for an incident beam of optical power P is given by [2]

$$\langle r \rangle = \frac{\eta P}{hv} \tag{14.6}$$

The production of electrons due to the incident photons at constant rate $\langle r \rangle$ is randomly distributed in time and obeys Poisson statistics, so that the probability of the production of m electrons in some measurement interval τ is given by

$$p(m, \tau) = (\langle r \rangle \tau)^m \frac{1}{m!} e^{-\langle r \rangle \tau} \tag{14.7}$$

The statistics involved with optical detection are very important in the determination of minimum detectable signal levels and, hence, the ultimate sensitivity of the sensors. At this point, however, we just note that the *electrical current is proportional to the optical power* incident on the detector:

$$i = \langle r \rangle e = \frac{\eta e P}{hv}, \tag{14.8}$$

where e is the charge of an electron. A change in input power ΔP (e.g., due to intensity modulation in a sensor) results in the output current Δi. Because power is proportional to squared current, the detector's electrical power output varies quadratically with input optical power, making it a "square-law" detector.

The voltage-to-current response of a typical photodiode is shown in Fig. 14.4B. If we attach a high-input-impedance voltmeter to the diode (corresponds to the case

when $i = 0$), we will observe that with increasing optical power, the voltage changes in a quite nonlinear fashion. In fact, variations are logarithmic. For the short-circuit conditions ($V = 0$), [i.e., when the diode is connected to a current-to-voltage converter (Fig. 5.10B of Chapter 5)], current varies linearly with the optical power. The current-to-voltage response of the photodiode is given by [3]

$$i = i_0(e^{eV/k_bT} - 1) - i_s, \tag{14.9}$$

where i_0 is a reverse "dark current" which is attributed to the thermal generation of electron–hole pairs, i_s is the current due to the detected optical signal, k_b is Boltzmann constant, and T is the absolute temperature. Combining Eqs. (14.8) and (14.9) yields

$$i = i_0(e^{eV/k_bT} - 1) - \frac{\eta e P}{h\nu}, \tag{14.10}$$

which is the overall characteristic of a photodiode. An efficiency of the direct conversion of optical power into electric power is quite low. Typically, it is in the range 5–10%; however, in 1992, it was reported that some experimental photocells were able to reach an efficiency as high as 25%. In sensor technologies, however, photocells are generally not used. Instead, an additional high-resistivity intrinsic layer is present between p and n types of the material, which is called a PIN photodiode (Fig. 14.5). The depth to which a photon can penetrate a photodiode is a function of its wavelength which is reflected in a spectral response of a sensor (Fig. 14.2).

In addition to very popular PIN diodes, several other types of photodiode are used for sensing light. In general, depending on the function and construction, all photodiodes may be classified as follows:

1. The *PN photodiodes* may include a SiO_2 layer on the outer surface (Fig. 14.6A). This yields a low-level dark current. To fabricate a high-speed version of the diode, the depletion layer is increased, thus reducing the junction capacitance (Fig. 14.6B). To make the diode more sensitive to ultraviolet (UV) light, a p layer can be made extra thin. A version of the planar diffusion type is a pnn^+ diode (Fig. 14.6C), which has a lower sensitivity to infrared and higher sensitivity at shorter wavelengths. This is due primarily to a thick layer of a low-resistance n^+ silicon to bring the nn^+ boundary closer to the depletion layer.

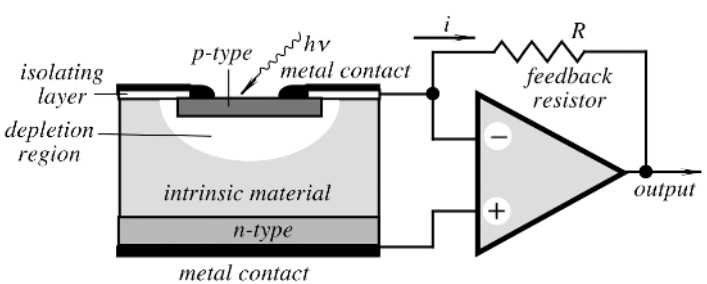


Fig. 14.5. Structure of a PIN photodiode connected to a current-to-voltage converter.

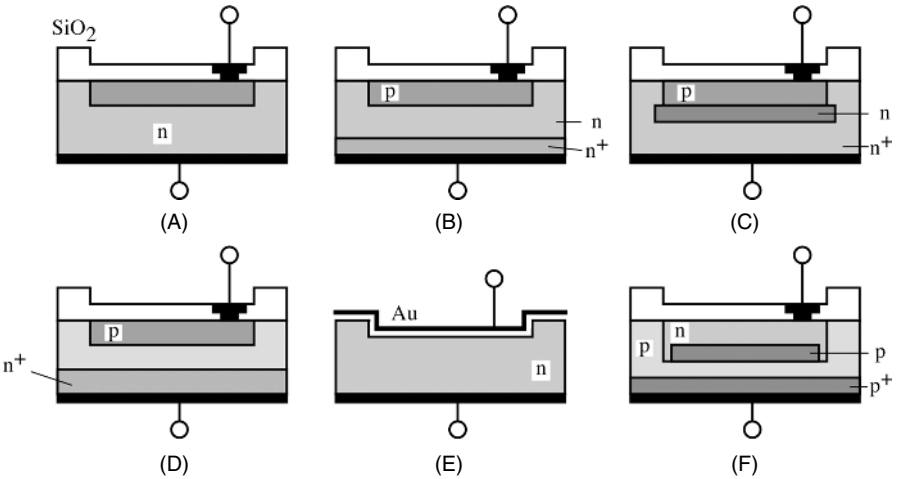


Fig. 14.6. Simplified structures of six types of photodiode.

2. The *PIN photodiodes* (Fig. 14.6D) are an improved version of low-capacitance planar diffusion diodes. The diode uses an extra high-resistance I layer between the p and n layers to improve the response time. These devices work even better with reversed bias, therefore, they are designed to have low leakage current high breakdown voltage.
3. The Schottky photodiodes (Fig. 14.6E) have a thin gold coating sputtered onto the n layer to form a Schottky p-n junction. Because the distance from the outer surface to the junction is small, the UV sensitivity is high.
4. The *avalanche photodiodes* (Fig. 14.6F) are named so because if a reverse bias is applied to the p-n junction and a high-intensity field is formed with the depletion layer, photon carriers will be accelerated by the field and collide with the atoms, producing the secondary carriers. In turn, the new carriers are accelerated again, resulting in the extremely fast avalanche-type increase in current. Therefore, these diodes work as amplifiers, making them useful for detecting extremely low levels of light.

There are two general operating modes for a photodiode: the *photoconductive* (PC) and the *photovoltaic* (PV). No bias voltage is applied for the photovoltaic mode. The result is that there is no dark current, so there is only thermal noise present. This allows much better sensitivities at low light levels. However, the speed response is worst due to an increase in C_j and responsivity to longer wavelengths is also reduced.

Figure 14.7A shows a photodiode connected in a PV mode. In this connection, the diode operates as a current-generating device which is represented in the equivalent circuit by a current source i_p (Fig. 14.7B). The load resistor R_b determines the voltage developed at the input of the amplifier and the slope of the load characteristic is proportional to that resistor (Fig. 14.7C).

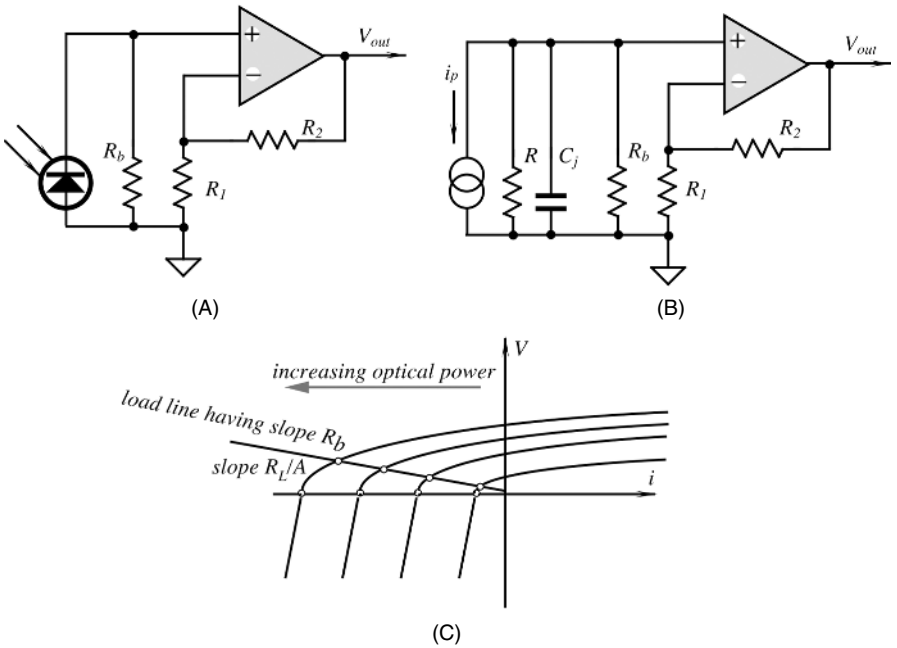


Fig. 14.7. Connection of a photodiode in a photovoltaic mode to a noninverting amplifier (A); the equivalent circuit (B); and a loading characteristic (C).

When using a photodiode in a photovoltaic mode, its large capacitance C_j may limit the speed response of the circuit. During the operation with a direct resistive load, as in Fig. 14.7A, a photodiode exhibits a bandwidth limited mainly by its internal capacitance C_j. Figure 14.7B models such a bandwidth limit. The photodiode acts primarily as a current source. A large resistance R and the diode capacitance shunt the source. The capacitance ranges from 2 to 20,000 pF depending, for the most part, on the diode area. In parallel with the shunt is the amplifier's input capacitance (not shown) which results in a combined input capacitance C. The diode resistance usually can be ignored, as it is much lower than the load resistance R_b. The net input network determines the input circuit response rolloff. The resulting input circuit response has a break frequency $f_1 = 1/2\pi R_L C$, and the response is [4]

$$V_{\text{out}} = \frac{-R_L i_p}{1 + j\frac{f}{f_1}}. \tag{14.11}$$

For a single-pole response, the circuit's 3-dB bandwidth equals the pole frequency. The expression reflects a typical gain-versus-bandwidth compromise. Increasing R_b gives a greater gain, but reduces f_1. From a circuit perspective, this compromise results from impressing the signal voltage on the circuit capacitances. The signal voltage appears across the input capacitance $C = C_j + C_{\text{OPAM}}$. To avoid the compromise, it

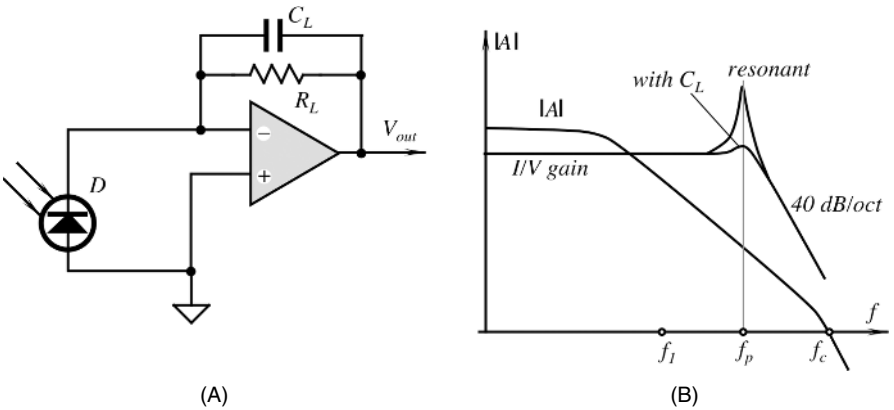


Fig. 14.8. Use of current-to-voltage converter (A) and the frequency characteristics (B).

is desirable to develop input voltage across the resistor and prevent it from charging the capacitances. This can be achieved by employing a current-to-voltage amplifier (I/V) as shown in Fig. 14.8A. The amplifier and its feedback resistor R_L translate the diode current into a buffered output voltage with excellent linearity. Added to the figure is a feedback capacitor C_L that provides a phase compensation. An ideal amplifier holds its two inputs at the same voltage (ground in the figure), thus the inverting input is called a virtual ground. The photodiode operates at zero voltage across its terminals, which improves the response linearity and prevents charging the diode capacitance. This is illustrated in Fig. 14.7C, where the load line virtually coincides with the current axis, because the line's slope is inversely proportional to the amplifier's open-loop gain A.

In practice, the amplifier's high, but finite, open-loop gain limits the performance by developing a small, albeit nonzero, voltage across the diode. Then, the break frequency is defined as

$$f_p = \frac{A}{2\pi R_L C} \approx A f_1, \tag{14.12}$$

where A is the open-loop gain of the amplifier. Therefore, the break frequency is increased by a factor A as compared with f_1. It should be noted that when the frequency increases, the gain, A, declines and the virtual load attached to the photodiode appears to be inductive. This results from the phase shift of gain A. Over most of the amplifier's useful frequency range, A has a phase lag of 90°. The 180° phase inversion by the amplifier converts this to a 90° phase lead, which is specific for the inductive impedance. This inductive load resonates with the capacitance of the input circuit at a frequency equal to f_p (Fig. 14.8B) and may result in an oscillating response (Fig. 14.9) or circuit instability. To restore stability, a compensating capacitor C_L is placed across the feedback resistor. The value of the capacitor can be found from

$$C_L = \frac{1}{2\pi R_L f_p} = \sqrt{C C_c}, \tag{14.13}$$

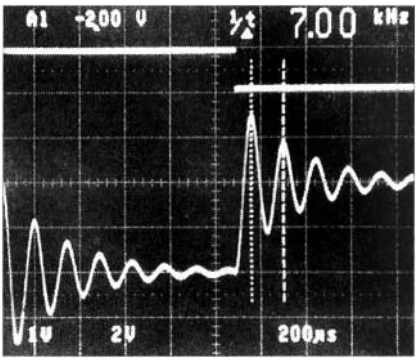


Fig. 14.9. Response of a photodiode with an uncompensated circuit. (Courtesy of Hamamatsu Photonics K.K.)

where $C_c = 1/(2\pi R_L f_c)$, and f_c is the unity-gain crossover frequency of the operational amplifier. The capacitor boosts the signal at the inverting input by shunting R_L at higher frequencies.

When using photodiodes for the detection of low-level light, the noise floor should be seriously considered. There are two main components of noise in a photodiode: shot noise and Johnson noise (see Section 5.9 of Chapter 5). In addition to the sensor, the amplifier's and auxiliary component noise also should be taken into account [see Eq. (5.75) of Chapter 5].

For the photoconductive (PC) operating mode, a reverse-bias voltage is applied to the photodiode. The result is a wider depletion region, lower junction capacitance C_j, lower series resistance, shorter rise time, and linear response in photocurrent over a wider range of light intensities. However, as the reverse bias is increased, the shot noise increases as well due to the increase in dark current. The PC mode circuit diagram is shown in Fig. 14.10A and the diode's load characteristic is in Fig. 14.10B. The reverse bias moves the load line into the third quadrant, where the response linearity is better than that for the PV mode (the second quadrant). The load lines

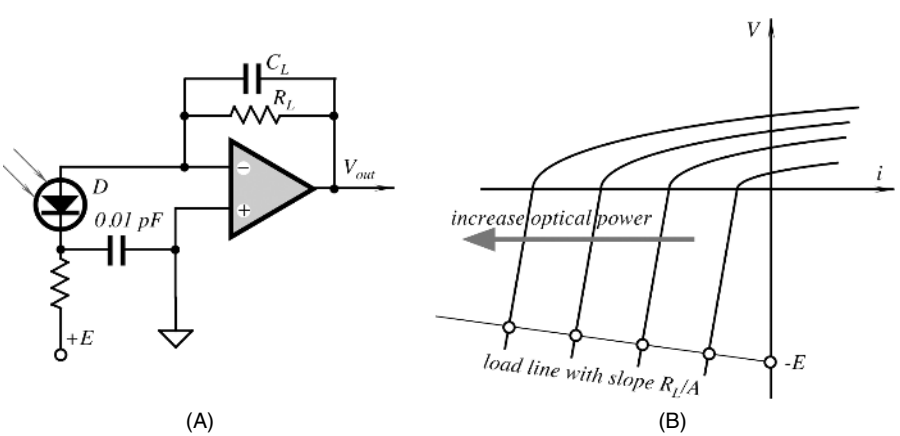


Fig. 14.10. Photoconductive operating mode: (A) a circuit diagram; (B) a load characteristic.

crosses the voltage axis at the point corresponding to the bias voltage E, and the slope is inversely proportional to the amplifier's open-loop gain A. The PC mode offers bandwidths to hundreds of megahertz, providing an accompanying increase in the signal-to noise ratio.

14.3 Phototransistor

A photodiode directly converts photons into charge carriers—specifically one electron and one hole (hole–electron pair) per a photon. Phototransistors can do the same, and in addition can provide current gain, resulting in a much higher sensitivity. The collector-base junction is a reverse-bias diode which functions as described earlier. If the transistor is connected into a circuit containing a battery, a photo-induced current flows through the loop, which includes the base–emitter region. This current is amplified by the transistor in the same manner as in a conventional transistor, resulting in a significant increase in the collector current.

The energy bands for the phototransistor are shown in Fig. 14.11. The photon-induced base current is returned to the collector through the emitter and the external circuitry. In so doing, electrons are supplied to the base region by the emitter, where they are pulled into the collector by the electric field. The sensitivity of a phototransistor is a function of the collector–base diode quantum efficiency and also of the dc current gain of the transistor. Therefore, the overall sensitivity is a function of collector current.

When subjected to varying ambient temperature, the collector current changes linearly with a positive slope of about 0.00667/°C. The magnitude of this temperature coefficient is primarily a result of the increase in current gain versus temperature, because the collector–base photocurrent temperature coefficient is only about 0.001/°C. The family of collector current versus collector voltage characteristics is very much

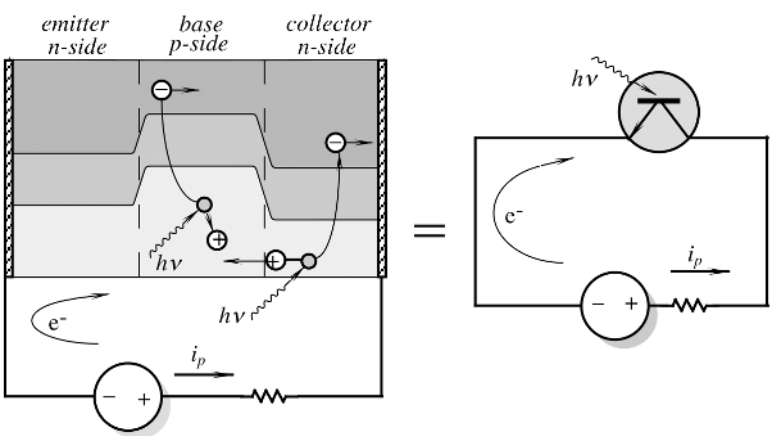


Fig. 14.11. Energy bands in a phototransistor.

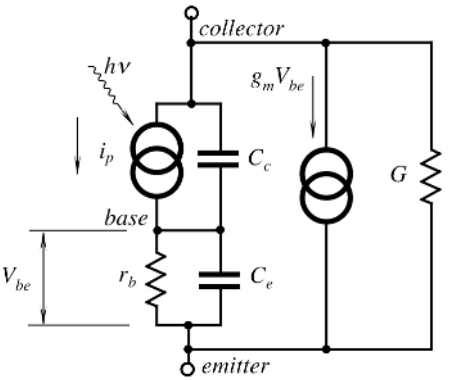


Fig. 14.12. An equivalent circuit of a phototransistor.

similar to that of a conventional transistor. This implies that circuits with phototransistors can be designed by using regular methods of transistor circuit techniques, except that its base should be used as an input of a photo-induced current which is supplied by its collector. Because the actual photogeneration of carriers occurs in the collector–base region, the larger the area of this region, the greater the number of carriers generated; thus, the phototransistor is so designed to offer a large area to impinging light. A phototransistor can be either a two-lead or a three-lead device. In the latter case, the base lead is available and the transistor may be used as a standard bipolar transistor with or without the additional capability of sensing light, thus giving a designer greater flexibility in circuit development. However, a two-lead device is the most popular as a dedicated photosensor.

When the base of the transistor is floating, it can be represented by an equivalent circuit shown in Fig. 14.12. Two capacitors C_c and C_e represent base–collector and base–emitter capacitances, which are the speed-limiting factors. Maximum frequency response of the phototransistor may be estimated from

$$f_1 \approx \frac{g_m}{2\pi C_e}, \tag{14.14}$$

where f_1 is the current–gain–bandwidth product and g_m is the transistor's forward transconductance.

Whenever a higher sensitivity of a photodetector is required, especially if a high response speed is not of a concern, an integrated Darlington detector is recommended. It is composed of a phototransistor whose emitter is coupled to the base of a bipolar transistor. Because a Darlington connection gives current gain equal to a product of current gains of two transistors, the circuit proves to be an efficient way of making a sensitive detector.

Spatial resolutions of both the light source and the detector must be seriously considered for many sensor applications. Whenever a higher efficiency of sensing is required, optical components come in handy. Let us, for instance, take a point light source which should be detected by a photodetector (Fig. 14.13A). According to Eq. (14.10), the sensor's output is proportional to the received photonic power, which, in

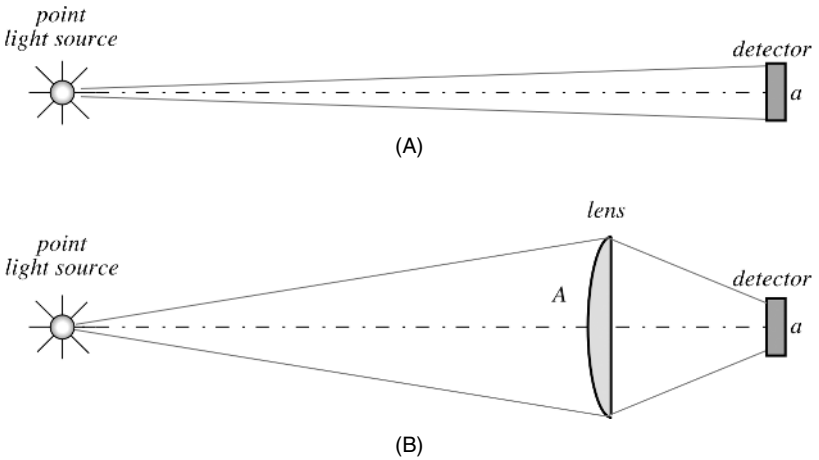


Fig. 14.13. Efficiency of a detector depends on its surface area a (A) or the area A of the focusing system (B).

turn, is proportional to the receiver's surface area. Figure 14.13B shows that the use of a focusing lens can dramatically increase the area. The efficiency of a single lens depends on its refractive index n. The overall improvement in the sensitivity can be estimated by employing Eqs. (4.5) and (4.8) of Chapter 4:

$$k \approx \frac{A}{a}\left[1 - 2\left(\frac{n-1}{n+1}\right)^2\right], \tag{14.15}$$

where A and a are effective areas of the lens and the sensing area of a photodetector, respectively. For glasses and most plastics operating in the visible and near-infrared spectral ranges, the equation can be simplified to

$$k \approx 0.92\frac{A}{a}. \tag{14.16}$$

It should be pointed out that the arbitrary placement of a lens may be more harmful than helpful; that is, a lens system must be carefully planed to be effective. For instance, many photodetectors have built-in lenses which are effective for parallel rays. If an additional lens is introduced in front of such a detector, it will create nonparallel rays at the input, resulting in the misalignment of the optical system and poor performance. Thus, whenever additional optical devices need to be employed, detector's own optical properties must be considered.

14.4 Photoresistors

As a photodiode, a photoresistor is a photoconductive device. The most common materials for its fabrication are cadmium sulfide[1] (CdS) and cadmium selenide (CdSe),

[1] Information on CdS photoresistors is courtesy of Hamamatsu Photonics K.K.

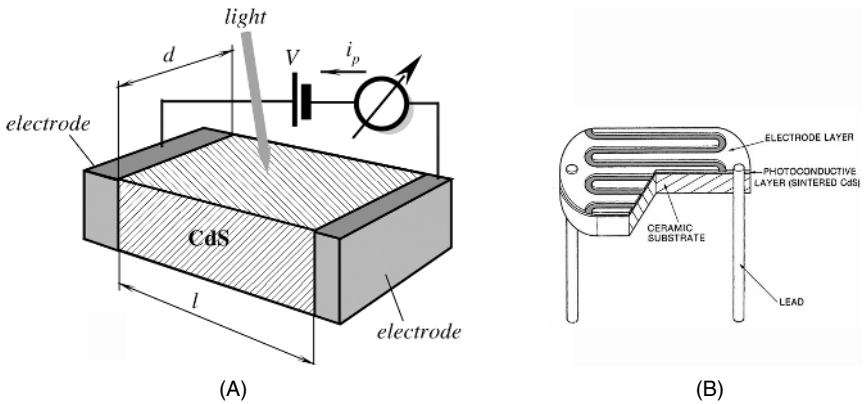


Fig. 14.14. Structure of a photoresistor (A) and a plastic-coated photoresistor having a serpentine shape (B).

which are semiconductors whose resistances change upon light entering the surface. For its operation, a photoresistor requires a power source because it does not generate photocurrent; a photoeffect is manifested in the change in the material's electrical resistance. Figure 14.14A shows a schematic diagram of a photoresistive cell. An electrode is set at each end of the photoconductor. In darkness, the resistance of the material is high. Hence, the applied voltage V results in a small dark current which is attributed to temperature effect. When light is incident on the surface, the current i_p flows.

The reason for the current increase is the following. Directly beneath the conduction band of the crystal is a donor level and there is an acceptor level above the valence band. In darkness, the electrons and holes in each level are almost crammed in place in the crystal, resulting in the high resistance of the semiconductor.

When light illuminates the photoconductive crystal, photons are absorbed, which results in the added-up energy in the valence band electrons. This moves them into the conduction band, creating free holes in the valence band, increasing the conductivity of the material. Because near the valence band there is a separate acceptor level that can capture free electrons not as easily as free holes, the recombination probability of the electrons and holes is reduced and the number of free electrons in the conduction band is high. Because CdS has a band gap of 2.41 eV, the absorption-edge wavelength is $\lambda = c/v \approx 515$ nm, which is in the visible spectral range. Hence, the CdS detects light shorter than 515-nm wavelengths. Other photoconductors have different absorption-edge wavelengths. For instance, CdS is most sensitive at the shorter-wavelength range, whereas Si and Ge are most efficient in the near infrared.

The conductance of a semiconductor is given by

$$\Delta\sigma = ef(\mu_n\tau_n + \mu_p\tau_p), \tag{14.17}$$

where μ_n and μ_p are the free-electron and hole movements (cm/V s), τ_n and τ_p are the free-electron and hole lives (s), e is the charge of an electron, and f is the number

of generated carriers per second per unit of volume. For a CdS cell, $\mu_n \tau_n \gg \mu_p \tau_p$; hence, conductance by free holes can be ignored. Then, the sensor becomes an n-type semiconductor. Thus,

$$\Delta\sigma = ef\mu_n\tau_n. \tag{14.18}$$

We can define the sensitivity b of the photoresistor through a number of electrons generated by one photon (until the carrier life span ends):

$$b = \frac{\tau_n}{t_t}, \tag{14.19}$$

where $t_t = l^2/V\mu_n$ is the transit time for the electron between the sensor's electrodes, l is the distance between the electrodes, and V is the applied voltage. Then, we arrive at

$$b = \frac{\mu_n \tau_n V}{l^2}. \tag{14.20}$$

For example, if $\mu_n = 300$ cm^2/V s, $\tau_n = 10^{-3}$ s, $l = 0.2$ mm, and $V = 1.2$ V, then the sensitivity is 900, which means that a single photon releases 900 electrons for conduction, making a photoresistor work as a photomultiplier. Indeed, a photoresistor is a very sensitive device.

It can be shown that for better sensitivity and lower cell resistance, the distance l between the electrodes should be reduced, and the width d of the sensor should be increased. This suggests that the sensor should be very short and very wide. For practical purposes, this is accomplished by fabricating a sensor in a serpentine shape (Fig. 14.14B) where the electrodes are connected to the leads.

Depending on the manufacturing process, the photoresistive cells can be divided into the sintered type, single-crystal type, and evaporated type. Of these, the sintered type offers high sensitivity and easier fabrication of large sensitive areas, which eventually translate into lower-cost devices. The fabrication of CdS cells consists of the following steps.

1. Highly pure CdS powder is mixed with appropriate impurities and a fusing agent.
2. The mixture is dissolved in water.
3. The solution in a form of paste is applied on the surface of a ceramic substrate and allowed to dry.
4. The ceramic subassemblies are sintered in a high-temperature oven to form a multicrystal structure. At this stage, a photoconductive layer is formed.
5. Electrode layers and leads (terminals) are attached.
6. The sensor is packaged into a plastic or metal housing with or without a window.

To tailor a spectral response of a photoresistor, the powder of step 1 can contain some variations; for instance, the addition of selenide or even the replacement of CdS for CdSe shifts the spectral response toward longer wavelengths (orange and red).

To illustrate, how photoresistors can be used, Fig. 14.15 shows two circuits. Circuit A shows an automatic light switch which turns lights on when illumination drops (the turn-off part of the circuit is not shown). Circuit B shows a beacon with a free-running multivibrator, which is enabled at darkness, when the resistance of a photoresistor becomes high.

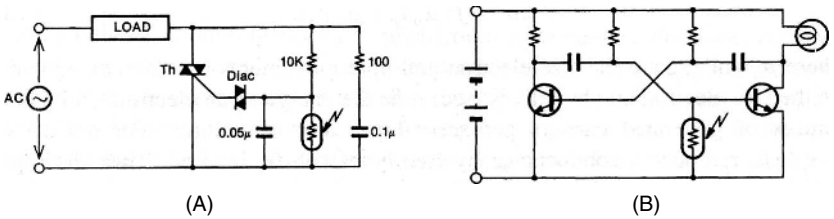


Fig. 14.15. Examples of photoresistor applications: (A) light switch and (B) beacon light. (Courtesy of Hamamatsu Photonics K.K.)

14.5 Cooled Detectors

For the measurement of objects emanating photons in the range of 2 eV or higher, quantum detectors having room temperature are generally used. For the smaller energies (longer wavelengths) narrower-band-gap semiconductors are required. However, even if a quantum detector has a sufficiently small energy band gap, at room temperatures its own intrinsic noise is much higher than a photoconductive signal. In other words, the detector will sense its own thermal radiation. The noise level is temperature dependent; therefore, when detecting long-wavelength photons, the signal-to-noise ratio may become so small that accurate measurement becomes impossible. This is the reason why, for the operation in the mid- and far-infrared spectral ranges, a detector not only should have a sufficiently narrow energy gap, but its temperature has to be lowered to the level where intrinsic noise is reduced to an acceptable level. Figure 14.16 shows typical spectral responses of some detectors with recommended operating temperatures. The operating principle of a cryogenically cooled detector is about the same as that of a photoresistor, except that it operates at far longer wavelengths at much lower temperatures. Thus, the sensor design becomes quite different. Depending on the required sensitivity and operating wavelength, the following crystals are typically used for this type of sensor: lead sulfide (PbS), indium arsenide (InAs), germanium (G), lead selenide (PbSe), and mercury–cadmium–telluride (HgCdTe).

Cooling shifts the responses to longer wavelengths and increases sensitivity. However, the response speeds of PbS and PbSe become slower with cooling. Methods of cooling include Dewar cooling using dry ice, or liquid nitrogen, liquid helium (Fig. 14.17), or thermoelectric coolers operating on the Peltier effect (see Section 3.9 of Chapter 3). As an example, Table 14.2 lists typical specifications for an MCT photoconductive detector. MCT stands for the mercury-cadmium-telluride type of a sensitive element.

Applications of the cryogenically cooled quantum detectors include the measurements of optical power over a broad spectral range, thermal temperature measurement and thermal imaging, detection of water content and gas analysis.

Figure 14.18 depicts gas absorption spectra for various molecules. Water strongly absorbs at 1.1, 1.4, 1.9, and 2.7 μm. Thus, to determine the moisture content, for example, in coal, the monochromatic light is projected on the test and reference samples. The reflected light is detected and the ratio is calculated for the absorption

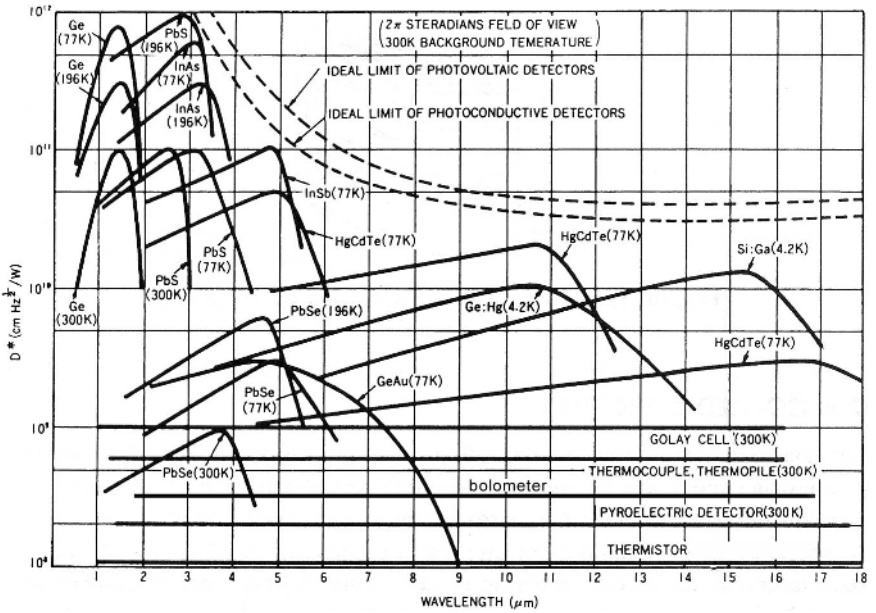


Fig. 14.16. Operating ranges for some infrared detectors.

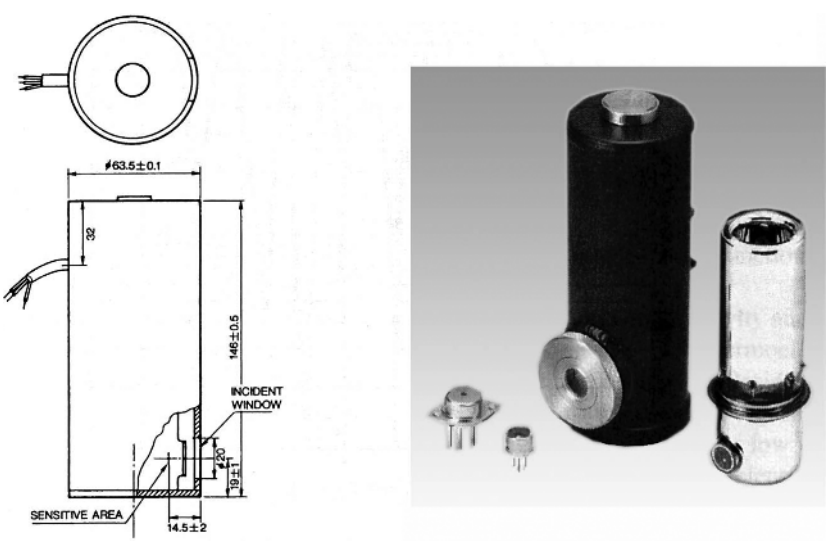


Fig. 14.17. Cryogenically cooled MCT quantum infrared detectors: (A) dimensional drawing of a Dewar type (in mm); (B) outside appearances of canned and Dewar detectors. (Courtesy of Hamamatsu Photonics K.K.)

Table 14.2. Typical Specifications for MCT Far-Infrared Detectors

Sensitive Area (mm)	Temperature (°C)	I_p (μm)	l_c (μm)	FOV (deg)	Dark Resist (kΩ)	Rise Time (μs)	Max. Current (mA)	D^* at l_p
1×1	−30	3.6	3.7	60	1	10	3	10^9
1×1	−196	15	16	60	20	1	40	3×10^9

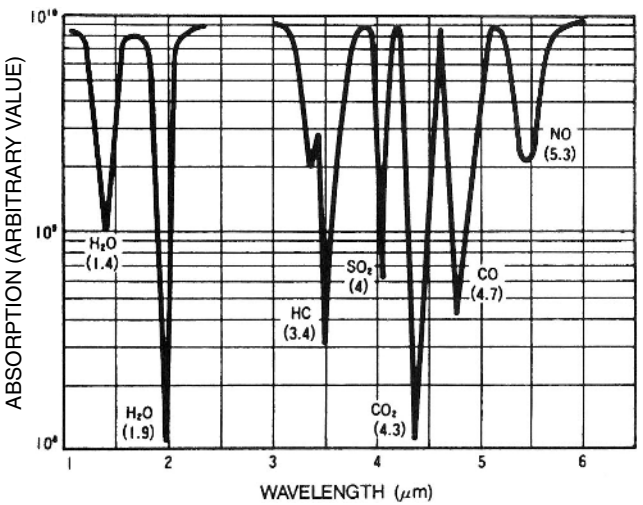


Fig. 14.18. Absorption spectra of gaseous molecules.

bands. The gas analyzer makes use of absorption in the infrared region of the spectrum. This allows us to measure gas density. Thus, it is possible to measure automobile exhaust gases (CO, HC, CO_2), emission control (CO, SO, NO_2), fuel leakage (CH_4, C_3H_2), and so forth.

14.6 Thermal Detectors

Thermal infrared detectors are primarily used for detecting infrared radiation in mid- and far-infrared spectral ranges and noncontact temperature measurements; these have been known for about 60 years in industry under the name *pyrometry* from the Greek word *pur* (fire). The respective thermometers are called radiation pyrometers. Today, noncontact methods of temperature measurement embrace a very broad range, including subzero temperatures, which are quite far away from that of flame. Therefore, it appears that *radiation thermometry* is a more appropriate term for this technology.

A typical infrared noncontact temperature sensor consists of the following:

1. A sensing element, which is responsive to electromagnetic radiation in the infrared wavelength range. The main requirements of the element are fast, predictable, and strong response to thermal radiation, and a good long-term stability.
2. A supporting structure to hold the sensing element and to expose it to the radiation. The structure should have low thermal conductivity to minimize heat loss.
3. A housing, which protects the sensing element from the environment. It usually should be hermetically sealed and often filled either with dry air or inert gas, such as argon or nitrogen.
4. A protective window which is impermeable to environmental factors and transparent in the wavelength of detection. The window may have surface coatings to improve transparency and to filter out undesirable portions of the spectrum.

Below the mid-infrared range, thermal detectors are much less sensitive than quantum detectors. Their operating principle is based on a sequential conversion of thermal radiation into heat and, then, conversion of heat level or heat flow into an electrical signal by employing conventional methods of heat detection. In principle, any temperature detector can be used for the detection of thermal radiation. However, according to Eq. (3.133) of Chapter 3, the infrared flux which is absorbed by a thermal detector is proportional to a geometry factor A which, for a uniform spatial distribution of radiation, is equal to the sensor's area. For instance, if a thermal radiation sensor at 25°C, having a 5-mm^2 surface area and ideal absorptivity, is placed inside a radiative cavity whose temperature is 100°C, the sensor will receive an initial radiative power of 3.25 mW. Depending on the sensor's thermal capacity, its temperature will rise until thermal equilibrium between the sensor and its environment occurs. It should be noted that, in practice, the sensing element's temperature never reaches that of an object. Unlike a hypothetical sensing element that is placed inside a radiative cavity, a real sensing element is rather poorly thermally coupled to the heat source. Although the element exchanges heat by radiation, a substantial portion of the heat is lost through a supporting structure and wires, as well as through gravitational convection and also through stray radiation. Thus, the equilibrium temperature is always somewhere in between the object's temperature and the initial temperature of the thermal detector.

All thermal radiation detectors can be divided into two classes: *passive* infrared (PIR) and *active* far-infrared (AFIR) detectors. Passive detectors absorb incoming radiation and convert it to heat, whereas active detectors generate heat from the excitation circuit.

14.6.1 Golay Cells

Golay cells are the broadband detectors of infrared radiation. They are extremely sensitive, but usually very delicate. The operating principle of the cell is based on the detection of a thermal expansion of gas trapped inside an enclosure. This is why these detectors sometimes are called thermopneumatic detectors. Figure 14.19 depicts an enclosed chamber having two membranes: the upper and lower. The upper membrane is coated with a heat absorber (gold black, e.g.) and the lower membrane has a mirror surface (coated with aluminum, e.g.).

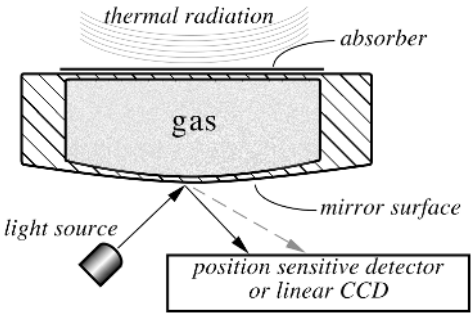


Fig. 14.19. Golay cell detector for mid- and far-infrared radiation.

The mirror is illuminated by a light source. The incident light beam is reflected from the mirror and impinges on a position-sensitive detector (PSD). The upper membrane is exposed to infrared radiation, which is absorbed by the coating and elevates the temperature of the membrane. This, in turn, warms up gas which is trapped inside the sensor. Gas expands and its pressure increases. The increase in the internal pressure deflects the lower membrane, which bulges out. A change in the mirror curvature modulates the direction of the reflected light beam. The reflected light impinges on the PSD at various locations, depending on the degree of bulging of the membrane and, therefore, on the intensity of the absorbed radiation. The entire sensor may be micromachined using modern MEMS technology (see Chapter 18). The degree of the lower membrane deflection alternatively may be measured by different methods [e.g., by using a Fabry–Perot (FP) interferometer; see Section 7.5 of Chapter 7].

14.6.2 Thermopile Sensors

Thermopiles belong to a class of PIR detectors. Their operating principle is the same as that of thermocouples. In effect, a thermopile is serially connected thermocouples. Originally, it was invented by Joule to increase the output signal of a thermoelectric sensor. He connected several thermocouples in a series and thermally joined their hot junctions. Currently, thermopiles have a different configuration. Their prime application is in the thermal detection of light in the mid- and far-infrared spectral ranges.

An equivalent schematic of a thermopile sensor is shown in Fig. 14.20A. The sensor consists of a frame having a relatively large thermal mass which is the place where the "cold" junctions are positioned. The frame may be thermally coupled with a reference temperature sensor or attached to a thermostat having a known temperature. The base supports a thin membrane whose thermal capacity and thermal conductivity are small. The membrane is the surface where the "hot" junctions are positioned. The words *hot* and *cold* are the remnants of traditional thermocouple jargon and are used here conditionally because the junctions in reality are rarely cold or hot.

The operating principle of a thermopile is the same as of any PIR detector. Infrared light is absorbed by or emanated from the membrane and changes its temperature. Because the membrane carries "hot" junctions, the temperature differential with respect to the "cold" junction generates thermoelectric voltage. The membrane temperature

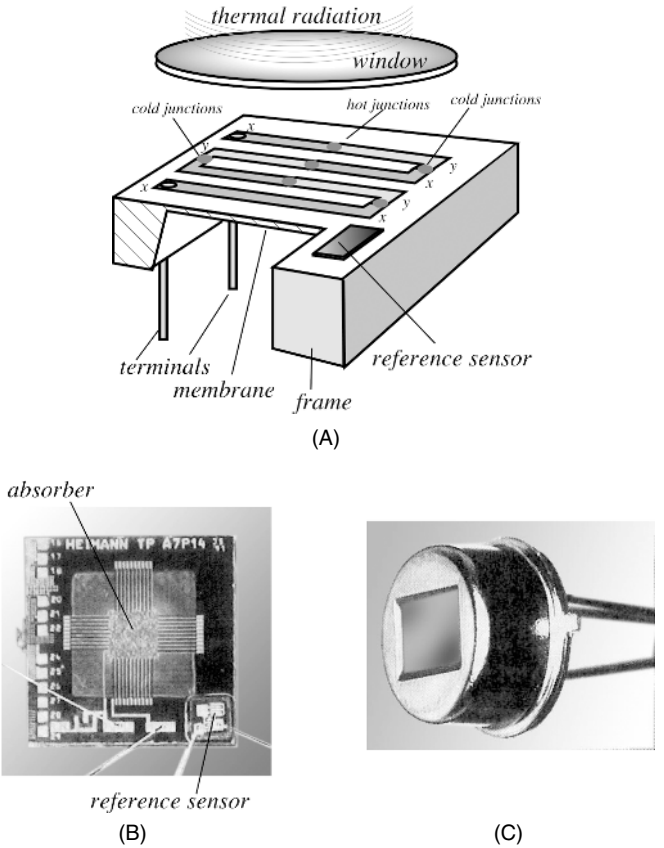


Fig. 14.20. Thermopile for detecting thermal radiation: (A) Equivalent schematic with a reference temperature sensor attached; x and y are different materials; (B) micromachined thermopile sensor; note the semiconductor reference temperature sensor on the silicon frame where the cold junctions are deposited and the absorptive coating on the hot junctions in the center of the membrane; (C) sensor in a TO-5 packaging.

increase depends on the thermal capacity, thermal conductivity, and intensity of the infrared light.

The best performance of a thermopile is characterized by high sensitivity and low noise, which may be achieved by the junction materials having a high thermoelectric coefficient α, low thermal conductivity, and low volume resistivity. In addition, the junction pairs should have thermoelectric coefficients of the opposite signs. This dictates the selection of the materials. Unfortunately, most of metals having low electrical resistivity (gold, copper, silver) have only very poor thermoelectric coefficients. The higher-electrical-resistivity metals (especially bismuth and antimony) possess high thermoelectric coefficients and they are the prime selection for designing thermopiles. By doping these materials with Se and Te, the thermoelectric coefficient has been improved up to 230 μV K^{-1} [5].

Table 14.3. Typical Specifications of a Thermopile.

Parameter	Value	Unit	Conditions
Sensitive area	0.5–2	mm^2	
Responsivity	50	V/W	6–14 μm, 500 K
Noise	30	$nV/\sqrt{Hz}$	25°C, rms
Equivalent resistance	50	kΩ	25°C
Thermal time constant	60	ms	
Temperature coefficient of resistivity	0.15	%/K	
Temperature coefficient of responsivity	−0.2	%/K	
Operating temperature	−20 to +80	°C	
Storage temperature	−40 to 100	°C	
Price	2–20	US$	

Methods of construction of metal junction thermopiles may differ to some extent, but all incorporate vacuum-deposition techniques and evaporation masks to apply the thermoelectric materials, such as bismuth and antimony. The number of junctions varies from 20 to several hundreds. The "hot" junctions are often coated with an absorber of thermal radiation. For example, they may be blackened (with gold black or organic paint) to improve their absorptivity of the infrared radiation.

A thermopile is a dc device whose output voltage follows its "hot" junction temperature quite well. A thermopile can be modeled as a thermal flux-controlled voltage source which is connected in series with a fixed resistor. The sensor is hermetically sealed in a metal can with a hard infrared transparent window, (e.g., silicon, germanium, or zinc selenide) (Fig. 14.20C). The output voltage V_s is nearly proportional to the incident radiation. The thermopile operating frequency limit is mainly determined by thermal capacity and thermal conductivity of the membrane, which are manifested through a thermal time constant. The sensor exhibits quite a low noise which is equal to the thermal noise of the sensor's equivalent resistance, (i.e., of 20–50 kΩ). Typical properties of a metal-type thermopile sensor are given in Table 14.3.

The output signal of a thermopile sensor depends on a temperature gradient between the source of the thermal radiation and the sensing surface. As a result, the transfer function of a thermopile is a three-dimensional surface whose shape is governed by the Stefan–Boltzmann law (see Fig. 2.1 of Chapter 2).

Currently, bismuth and antimony are being replaced by silicon thermopiles. These thermopiles are more efficient and reliable [6]. Table A.11 in the Appendix lists thermoelectric coefficients for selected elements. It is seen that the coefficients for crystalline and polycrystalline silicon are very large and the volume resistivity is relatively low. The advantage of using silicon is in the possibility of employing standard integrated circuit (IC) processes which result in a significant cost reduction. The resistivity and the thermoelectric coefficients can be adjusted by the doping concentration. However, the resistivity increases much faster and the doping concentration has to be carefully optimized for the high sensitivity–low noise ratios.

Figure 14.20B shows a semiconductor thermopile sensor produced by Perkin-Elmer Optoelectronics (Wiesbaden, Germany). It is fabricated by employing a micro-

machining (MEMS) technology. The central part of the silicon substrate is removed by means of anisotropic etching from the back, leaving only about a 1-μm thin sandwich layer (membrane) of SiO_2–Si_3N_4 on top, which has a low thermal conductivity. Onto this membrane thin conductors of two different thermoelectric materials (polysilicon and aluminum) are deposited. This allowed the production of sensors with a negligible temperature coefficient of sensitivity, which is an important factor for operation over broad ambient temperatures ranges.

14.6.3 Pyroelectric Sensors

Pyroelectric sensors belong to a class of PIR detectors. A typical construction of a solid-state pyroelectric sensor is shown in Fig. 14.21A. It is housed in a metal TO-5 or TO-39 can for better shielding and is protected from the environment by a silicon or any other appropriate window. The inner space of the can is often filled with dry air or nitrogen. Usually, two sensing elements are oppositely, serially, or in parallel connected for better compensation of rapid thermal changes and mechanical stresses resulting from acoustical noise and vibrations. Sometimes, one of the elements is coated with heat-absorbing paint or gold black and the other element is shielded from radiation and gold plated for better reflectivity. Alternatively, the element is given nichrome electrodes. Nichrome has high emissivity and thus serves a dual purpose: to collect electric charge and to absorb thermal radiation. For applications in PIR motion detectors, both pyroelectric elements are exposed to the window.

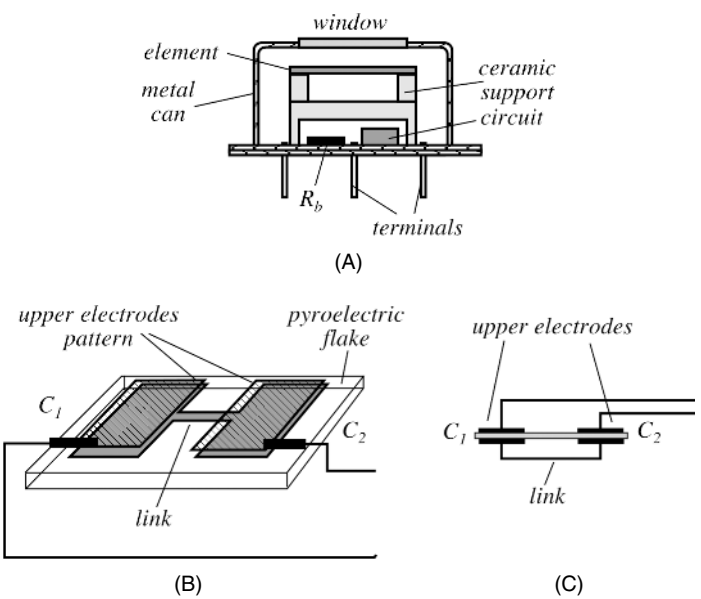


Fig. 14.21. A dual-pyroelectric sensor: (A) structure of a sensor in a metal can; (B) metal electrodes are deposited on the opposite sides of a material; (C) equivalent circuit of a dual element.

A dual element is often fabricated from a single flake of a crystalline material (Fig. 14.21B). The metallized pattern on both sides of the flake form two serially connected capacitors C_1 and C_2. Figure 14.21C shows an equivalent circuit of a dual-pyroelectric element. This design has the benefit of a good balance of both elements, thus resulting in a better rejection of common-mode interferences. Note that the sensing element exists only between the opposite electrodes and the portion of the flake that is not covered by the electrodes, is not participating in the generation of a useful signal. A major problem in the design of pyroelectric detectors is in their sensitivity to mechanical stress and vibrations. All pyroelectrics are also piezoelectrics; therefore, although sensitive to thermal radiation, the pyroelectric sensors are susceptible to interferences which are called "microphonics" sometimes. For better noise rejection, the crystalline element must be mechanically decoupled from the outside, especially from the terminals and the metal can.

A pyroelectric element (a crystal flake plus two opposite electrodes) can be modeled by a capacitor connected in parallel with a leakage resistor. The value of that resistor is on the order of $10^{12} - 10^{14}\,\Omega$. In practice, the sensor is connected to a circuit which contains a bias resistor R_b and an impedance converter ("circuit" in Fig. 14.21A). The converter may be either a voltage follower (e.g., JFET transistor) or a current-to-voltage converter. The voltage follower (Fig. 14.22A) converts the high output impedance of the sensor (capacitance C in parallel with a bias resistance R_b) into the output resistance of the follower which, in this example, is determined by the transistor's transconductance in parallel with 47 kΩ. The advantage of this circuit is in its simplicity, low cost, and low noise. A single JFET follower is the most cost-effective and simple; however, it suffers from two major drawbacks. The first is the dependence of its speed response on the so-called *electrical time constant*, which is a product of the sensor's capacitance C and the bias resistor R_b:

$$\tau_e = C R_b. \tag{14.21}$$

For example, a typical dual sensor may have $C = 40$ pF and $R_b = 50$ GΩ, which yield $\tau_e = 2$ s, corresponding to a first-order frequency response with the upper cutoff frequency at the 3-dB level equal to about 0.08 Hz—a very low frequency indeed.

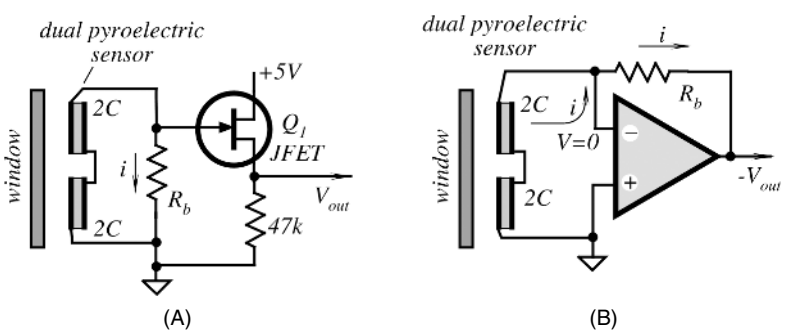


Fig. 14.22. Impedance converters for pyroelectric sensors: (A) voltage follower with JFET; (B) current-to-voltage converter with operational amplifier.

This makes the voltage follower suitable only for limited applications, where speed response in not too important. An example is the detection of movement of people (see Chapter 6). The second drawback of the circuit is a large offset voltage across the output resistor. This voltage depends on the type of the transistor and is temperature dependent. Thus, the output V_{out} is the sum of two voltages: the offset voltage, which can be as large as several volts, and the alternate pyroelectric voltage, which may be on the order of millivolts.

A more efficient, but more expensive, circuit for a pyroelectric sensor is an I/V (current-to-voltage) converter (Fig. 14.22B). Its advantage is its faster response and insensitivity to the capacitance of the sensor element. The sensor is connected to an inverting input of the operational amplifier which possesses properties of the so-called virtual ground (similar circuits are shown in Figs. 14.5, 14.8, and 14.10); that is, the voltage in the inverting input is constant and almost equal to that of a noninverting input, which is grounded in this circuit. Thus, the voltage across the sensor is forced by the feedback to stay near zero. The output voltage follows the shape of the electric current (a flow of charges) generated by the sensor (Fig. 3.28 of Chapter 3). The circuit should employ an operational amplifier with a very low bias current (on the order of 1 pA). There are three major advantages in using this circuit: a fast response, insensitivity to the capacitance of the sensor, and a low-output offset voltage. However, being a broad-bandwidth circuit, a current-to-voltage converter may suffer from higher noise.

At very low frequencies, both circuits, the JFET and I/V converter, transform pyroelectric current i_p into output voltage. According to Ohm's law,

$$V_{\text{out}} = i_p R_b. \tag{14.22}$$

For instance, for the pyroelectric current of 10 pA (10^{-11} A) and the bias resistor of $5 \times 10^{10}\,\Omega$ (50 GΩ), the output voltage is 500 mV. Either the JFET transistor or operational amplifier must have low input bias currents (I_B) over the entire operating temperature range. The CMOS (OPAMs) are generally preferable, as their bias currents are on the order of 1 pA.

It should be noted that the above-described circuits (Fig. 14.23) produce output signals of quite different shapes. The voltage follower's output voltage is a repetition of voltage across the element and R_b (Fig. 14.24A). It is characterized by two slopes: the leading slope having an electrical time constant $\tau_e = C R_b$, and the decaying slope having thermal time constant τ_T. Voltage across the element in the current-to-voltage converter is essentially zero and, contrary to the follower, the input impedance of the converter is low. In other words, whereas the voltage follower acts as a voltmeter, the current-to-voltage converter acts as an ampermeter. The leading edge of its output voltage is fast (determined by a stray capacitance across R_b) and the decaying slope is characterized by τ_T. Thus, the converter's output voltage repeats the shape of the sensor's pyroelectric current (Fig. 14.23B).

A fabrication of gigaohm-range resistors is not a trivial task. High-quality bias resistors must have good environmental stability, low temperature coefficient of resistance (TCR), and low voltage coefficient of resistance (VCR). The VCR is defined as

$$\xi = \frac{R_1 - R_{0.1}}{R_{0.1}} \times 100\%, \tag{14.23}$$

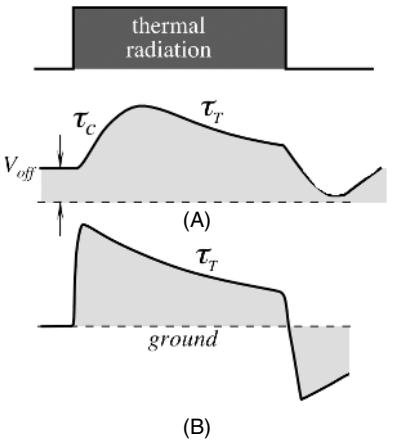


Fig. 14.23. Output signals of the voltage follower (A) and current-to-voltage converter (B) in response to a step function of a thermal radiation.

where R_1 and $R_{0.1}$ are the resistances measured respectively at 1 and 0.1 V. Usually, the VCR is negative; that is, the resistance value drops with an increase in voltage across the resistor (Fig. 14.24A). Since the pyroelectric sensor's output is proportional to the product of the pyroelectric current and the bias resistor, the VCR results in the nonlinearity of an overall transfer function of the sensor plus circuit. A high-impedance resistor is fabricated by depositing a thin layer of a semiconductive ink on a ceramic (alumina) substrate, firing it in a furnace and subsequently covering the surface with a protective coating. A high-quality, relatively thick (at least 50 μm thick) hydrophobic coating is very important for protection against moisture, because even a small amount of water molecules may cause oxidation of the semiconductive layer. This causes a substantial increase in the resistance and poor long-term stability. A typical design of a high-impedance resistor is shown in Fig. 14.24B.

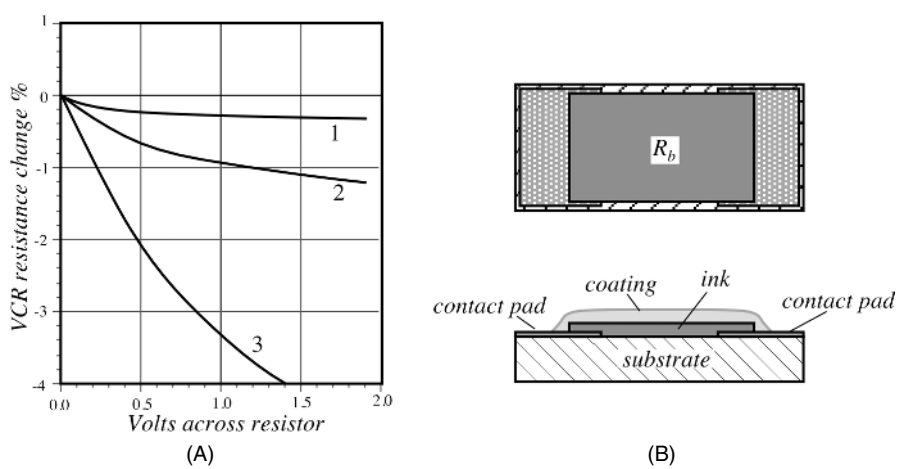


Fig. 14.24. High-impedance resistor: (A) VCRs for three different types of the resistor; (B) structure of a resistor on an alumina substrate.

In applications, where high accuracy is not required, such as thermal motion detection, the bias resistor can be replaced with one or two zero-biased parallel-opposite connected silicon diodes.

For the detection of thermal radiation, a distinction exists between two cases in which completely different demands have to be met with respect to the pyroelectric material and its thermal coupling to the environment [7]:

1. *Fast* sensors detect radiation of high intensity but very short duration (nanoseconds) of laser pulses, with a high repetition on the order of 1 MHz. The sensors are usually fabricated from single-crystal pyroelectrics, such as lithium tantalate ($LiTaO_3$) or triglycinesulfate (TGS). This assures a high linearity of response. Usually, the materials are bonded to a heat sink.
2. *Sensitive* sensors detect thermal radiation of low intensity, but, with a relatively low rate of change. Examples are infrared thermometry and motion detection [8–10]. These sensors are characterized by a sharp temperature rise in the field of radiation. This generally requires a good thermal coupling with a heat source. Optical devices, such as focusing lenses and waveguides, are generally employed. A heat transfer to the environment (sensor's housing) must be minimized. If well designed, such a sensor can have a sensitivity approaching that of a cryogenically cooled quantum detector [7]. Commercial pyroelectric sensors are implemented on the basis of single crystals, such as TGS and LiTaO3, or lead–zirconate–titanate (PZT) ceramics. Polyvinylidene fluoride (PVDF) film is also occasionally used because of its high-speed response and good lateral resolution.

14.6.4 Bolometers

Bolometers are miniature resistive temperature detectors (RTDs) or thermistors (see Section 16.1.3 of Chapter 16) or other temperature-sensitive resistors which are mainly used for measuring root-mean-square (r.m.s.) values of electromagnetic radiation over a very broad spectral range from mid-infrared to microwaves. Applications include infrared temperature detection and imaging, measurements of local fields of high power, the testing of microwave devices, radio-frequency (RF) antenna beam profiling, testing of high-power microwave weapons, monitoring of medical microwave heating, and others. The operating principle is based on a fundamental relationship between the absorbed electromagnetic signal and dissipated power [11]. The conversion steps in a bolometer are as follows:

1. An ohmic resistor is exposed to electromagnetic radiation. The radiation is absorbed by the resistor and converted into heat.
2. The heat elevates the resistor's temperature above the ambient.
3. The temperature increase reduces the bolometer ohmic resistance.

A temperature increase is a representation of the electromagnetic power. Naturally, this temperature differential can be measured by any suitable method. These methods are covered in Chapter 16. Here, we briefly outline the most common methods of bolometer fabrications which evolved quite dramatically since Langley first invented a bolometer over 100 years ago.

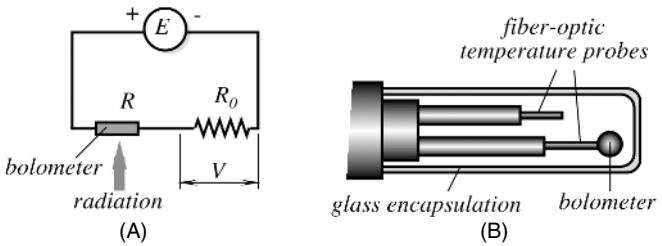


Fig. 14.25. Equivalent circuit of electrically biased bolometer (A) and a design of an optical bolometer (B).

A basic circuit for the voltage biased bolometer application is shown in Fig. 14.25A. It consists of a bolometer (a temperature-sensitive resistor) having resistance R, a stable reference resistor R_0, and a bias voltage source E. The voltage V across R_0 is the output signal of the circuit. It has the highest value when both resistors are equal. The sensitivity of the bolometer to the incoming electromagnetic (EM) radiation can be defined as [12]

$$\beta_v = \frac{\alpha \varepsilon Z_T E}{4\sqrt{1+(\omega\tau)^2}}, \tag{14.24}$$

where $\alpha = (dR/dT)/R$ is the TCR of the bolometer, ε is the surface emissivity, Z_T is the bolometer thermal resistance, which depends on its design and the supporting structure, τ is the thermal time constant, which depends on Z_T and the bolometer's thermal capacity, and ω is the frequency.

Because the bolometer's temperature increase, ΔT is

$$\Delta T = T - T_0 \approx P_E Z_T = \frac{E^2}{4R} Z_T, \tag{14.25}$$

and the resistance of a RTD bolometer can be represented by a simplification Eq. (16.14) of Chapter 16,

$$R = R_0(1 + \alpha_0 \Delta T), \tag{14.26}$$

Eq. (14.24) can be rewritten as

$$\beta_V = \frac{1}{2}\varepsilon\alpha \sqrt{\frac{R_0 Z_T \Delta T}{(1+\alpha_0\Delta T)\left[1+(\omega\tau)^2\right]}} \tag{14.27}$$

Therefore, to improve the bolometer's responsivity, its electrical resistance and thermal impedance should be increased.

The bolometers were traditionally fabricated as miniature thermistors, suspended by tiny wires. Another popular method of bolometer fabrication is the use of metal film depositions [12,13], usually of nichrome. In many modern bolometers, a thermoresistive thin-film material is deposited on the surface of a micromachined silicon or a glass membrane which is supported by a silicon frame. This approach gains popularity with the increased demand for the focal-plane array (FPA) sensors that are required for the thermal imaging. When an application does not need a high

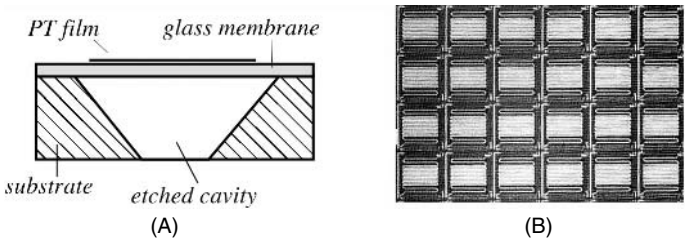


Fig. 14.26. Platinum-film bolometer: (A) glass membrane over the etched cavity; (B) array of bolometers.

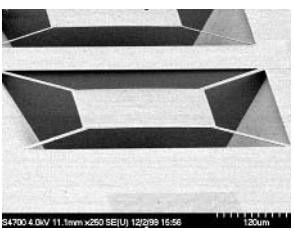


Fig. 14.27. Germanium-film bolometer floating over the silicon cavity. (Courtesy of Professor J. Shie.)

sensitivity and where the cost of fabrication is a critical factor, a platinum-film bolometer is an attractive choice. Platinum has a small but predictive temperature coefficient of resistivity.

The platinum film (having thickness of about 500 Å) is deposited and photolithographically patterned over the thin glass membrane. The membrane is supported in the cavity etched in silicon by tiny extended leads. Thus, the membrane plate is virtually floating over the V-grooved cavity in the Si substrate. This helps to dramatically minimize its thermal coupling with the substrate. Figure 14.26B shows a microphotograph of an array of the Pt bolometers used for the thermal imaging.

In addition to platinum, many other materials may be used as temperature-sensitive resistors, (e.g., polysilicon, germanium, TaNO, and others). An important issue when selecting a particular material is its compatibility with a standard CMOS process so that a full monolithic device can be fabricated on a single silicon chip, including the interface electronic circuit. Thus, polysilicon is an attractive choice, along with the deposition of germanium films (Fig. 14.27).

As follows from Eq. (14.27), one of the critical issues which always must be resolved when designing a bolometer (or any other accurate temperature sensor, for that matter) is to assure good thermal insulation of the sensing element from a supporting structure, connecting wires, and interface electronics. Otherwise, heat loss from the element may result in large errors and reduced sensitivity. One method for achieving this is to completely eliminate any metal conductors and to measure the temperature of the bolometer by using a fiber-optic technique, as has been implemented in the *E*-field probe fabricated by Luxtron (Mountain View, CA; (U.S. patent 4,816,634). In the design (Fig. 14.25B), a miniature bolometer is suspended in the end of an optical probe and its temperature is measured by a fluoroptic temperature sensor (see

Section 16.4.1 of Chapter 16) and another similar optical sensor measures ambient temperature to calculate ΔT.

14.6.5 Active Far-Infrared Sensors

In the active far-infrared (AFIR) sensor, a process of measuring thermal radiation flux is different from the previously described passive (PIR) detectors. Contrary to a PIR sensing element, whose temperature depends on both the ambient and object's temperatures, the AFIR sensor's surface is actively controlled by a special circuit to have a defined temperature T_s, which, in most applications, is maintained constant during an entire measurement process. To control the sensor's surface temperature, electric power P is provided by a control (or excitation) circuit (Fig. 14.28A). To regulate T_s, the circuit measures the element's surface temperature and compares it with an internal reference. Obviously, the incoming power maintains T_s higher than ambient. In some applications, T_s may be selected higher than the highest temperature of the object; however, in most cases, just several tenths of a degree Celsius above the ambient is sufficient. Because the element's temperature is above ambient, the sensing element loses thermal energy *toward* its surroundings, rather than passively absorbs it, as in a PIR detector. Part of the heat loss is in the form of a thermal conduction, part is a thermal convection, and another part is thermal radiation. That third part is the one which has to be measured. Unlike the conductive and convective heat transfer, which is always directed out of the sensing element (because it is warmer than ambient), the radiative heat transfer may go in either direction, depending on the temperature of the object. Of course, the radiative flux is governed by the fundamental Eq. (3.138) of Chapter 3, which is known as the Stefan–Boltzmann law.

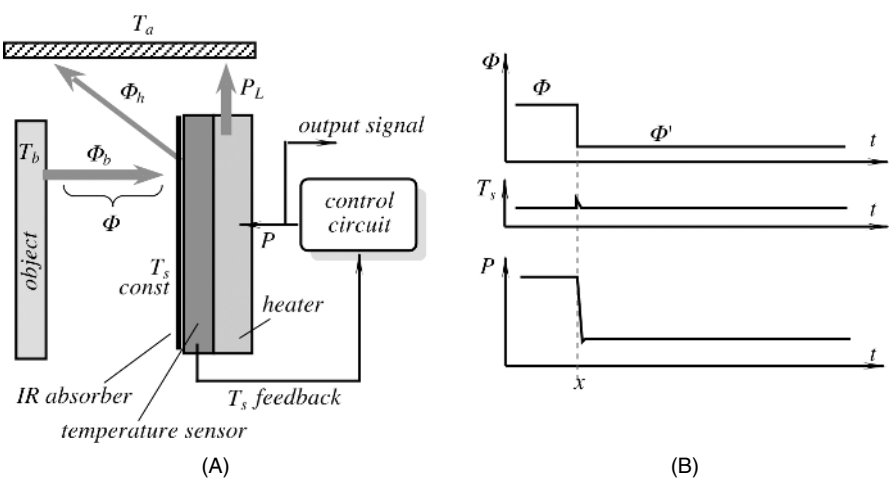


Fig. 14.28. The AFIR element radiates thermal flux Φ_η toward its housing and absorbs flux Φ_b from the object (A); timing diagrams for radiative flux, surface temperature, and supplied power (B).

Some of the radiation power goes out of the element to the sensor's housing, while some come from the object (or goes to the object). What is essential is that the net thermal flow (coductive+convective+radiative) always must come out of the sensor (e.g., it must have a negative sign).

If the AFIR element is provided with a cooling element (e.g., a thermoelectric device operating on Peltier effect[2], T_s may be maintained at or below ambient. However, from practical standpoint, it is easier to warm the element up rather than to cool it down. In the following, we discuss the AFIR sensors where the surface is warmed up either by an additional heating element or due to a self-heating effect in a temperature sensor [8,14–16].

Dynamically, the temperature T_s of any thermal element, either active or passive, in general terms may be described by the first-order differential equation

$$cm\frac{dT_s}{dt} = P - P_L - \Phi, \tag{14.28}$$

where P is the power supplied to the element from a power supply or an excitation circuit (if any), P_L is a nonradiative thermal loss which is attributed to thermal conduction and convection, m and c are the sensor's mass and specific heat, respectively, and $\Phi = \Phi_\eta + \Phi_b$ is the net radiative thermal flux. We select a positive sign for power P when it is directed toward the element.

In the PIR detector, for instance, in the thermopile or pyroelectric, no external power is supplied ($P = 0$), hence, the speed response depends only on the sensor's thermal capacity and heat loss and is characterized by a thermal time constant τ_T. In the AFIR element, after a warmup period, the control circuit forces the element's surface temperature T_s to stay constant, which means

$$\frac{dT_s}{dt} = 0, \tag{14.29}$$

and Eq. (14.28) becomes algebraic:

$$P = P_L + \Phi. \tag{14.30}$$

Contrary to PIR sensors, the AFIR detector acts as an "infinite" heat source. It follows from the above that under idealized conditions, its response does not depend on thermal mass and is not a function of time. If the control circuit is highly efficient, because P_L is constant at given ambient conditions, electronically supplied power P should track changes in the radiated flux Φ with high fidelity. A magnitude of that power may be used as the sensor's output signal. Equation (14.30) predicts that an AFIR element, in theory, is a much faster device if compared with the PIR. The efficiency of the AFIR detector is a function of both its design and the control circuit. Nonradiative loss P_L is a function of ambient temperature T_a and a loss factor α_s:

$$P_L = \alpha_s(T_s - T_a). \tag{14.31}$$

To generate heat in the AFIR sensor, it may be provided with a heating element having electrical resistance R. During the operation, electric power dissipated by the heating element is a function of voltage V across that resistance:

[2] See Section 3.9 of Chapter 3.

$$P = V^2/R. \tag{14.32}$$

Let us assume that the AFIR sensor is used in a radiation thermometer. Its output signal should be representative of the object's temperature T_b that is to be measured. Substituting Eq. (3.138) of Chapter 3 and Eqs. (14.31) and (14.32) into Eq. (14.30), assuming that $T = T_b$ and $T_s > T_a$, after simple manipulations the object's temperature may be presented as a function of voltage V across the heating element:

$$T_b = \sqrt[4]{T_s^4 - \frac{1}{A\sigma\varepsilon_s\varepsilon_b}\left[\frac{V^2}{R} - \alpha_s(T_s - T_a)\right]}. \tag{14.33}$$

The coefficient α_s is the thermal conductivity from the AFIR detector to the environment (housing).

Contrary to a PIR detector, an AFIR sensor is active and can generate a signal only when it works in concert with a control circuit. A control circuit must include the following essential components: a reference to the preset temperature, an error amplifier, and a driver stage. In addition, it may include an RC network for correcting a loop response function and for stabilizing its operation, otherwise an entire system may be prone to oscillations [17].

It may be noted that an AFIR sensor along with its control circuit is a direct converter of thermal radiative power into electric voltage and is quite an efficient one. Its typical responsivity is in the range of 3000 V/W, which is much higher compared to a thermopile whose typical responsivity is in the range of 100 V/W. An efficient way to fabricate an AFIR sensor would be by MEMS technology. In fact, an AFIR sensor is a close relative of a bolometer as described in the previous section. It just needs to be provided with a heater that can be deposited beneath the bolometer temperature-sensing element.

14.7 Gas Flame Detectors

Detection of a gas flame is very important for security and fire prevention systems. In many respects, it is a more sensitive way to detect fire than a smoke detector, especially outdoors, where smoke concentration may not reach a threshold level for alarm triggering.

To detect burning gas, it is possible to use a unique feature of the flame: A noticeable portion of its optical spectrum is located in the UV spectral range (Fig. 14.29). After passing through the atmosphere, sunlight loses a large portion of its UV spectrum located below 250 nm, whereas a gas flame contains UV components down to 180 nm. This makes it possible to design a narrow-bandwidth element for the UV spectral range which is selectively sensitive to flame and not sensitive to sunlight or electric lights.

An example of such a device is shown in Fig. 14.30A. The element is a UV detector that makes use of a photoelectric effect in metals along with the gas multiplication effect (see Chapter 14). The detector is a rare-gas-filled tube. The UV-transparent

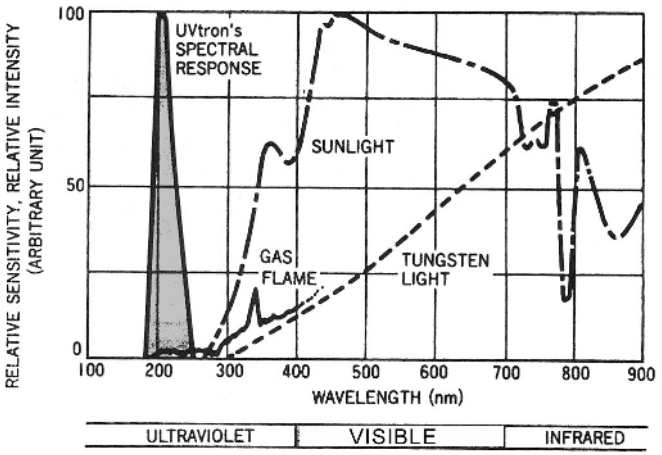


Fig. 14.29. Electromagnetic spectra of various sources. (Courtesy of Hamamatsu Photonics K.K.)

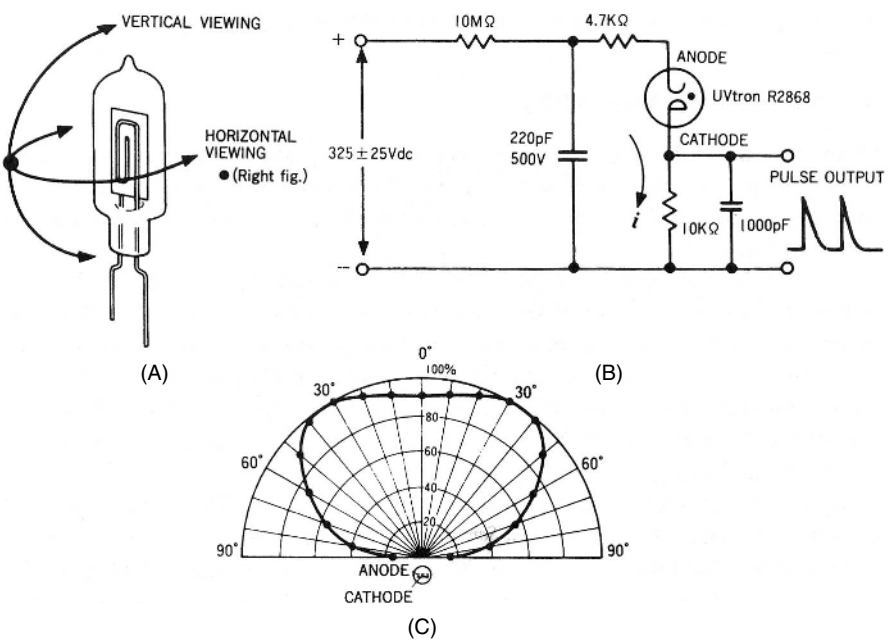


Fig. 14.30. UV flame detector. (A) a glass-filled tube; (B) angle of view in horizontal plane; (C) recommended operating circuit. (Courtesy of Hamamatsu Photonics K.K.)

housing assures wide angles of view in both horizontal and vertical planes (Fig. 14.30C). The device needs high voltage for operation, and under normal conditions, it is not electrically conductive. Upon being exposed to a flame, the high-energy UV photons strike the cathode, releasing free electrons to the gas-filled tube interior. Gas atoms receive an energy burst from the emitted electrons, which results in gas luminescence in the UV spectral range. This, in turn, cause more electrons to be emitted, which cause more UV luminescence. Thus, the element develops a fast avalanche-type electron multiplication, making the anode–cathode region electrically conductive. Hence, upon being exposed to a gas flame, the element works as a current switch, producing a strong positive voltage spike at its output (Fig. 14.30B). It follows from the above description that the element generates UV radiation in response to flame detection. Albeit being of a low intensity, the UV does not present harm to people; however, it may lead to cross-talk between similar neighboring sensors.

References

1. Chappell, A., ed. *Optoelectronics: Theory and Practice*. McGraw-Hill, New York, 1978.
2. Spillman, W. B., Jr. Optical detectors. In: *Fiber Optic Sensors*, E. Udd, ed. John Wiley & Sons, New York, 1991, pp. 69–97.
3. Verdeyen, J. T. *Laser Electronics*, Prentice-Hall, Englewood Clifs, NJ, 1981.
4. Graeme, J. Phase compensation optimizes photodiode bandwidth. Electronic Design News *(EDN)*, pp. 177–183, 1992.
5. Völklein, A. Wiegand A., and Baier, V. *Sensors Actuators A* 29, 87–91, 1991.
6. Schieferdecker, J., Quad, R., Holzenkämpfer, E., and Schulze, M. Infrared thermopile sensors with high sensitivity and very low temperature coefficient. *Sensors Actuators A* 46–47, 422–427, 1995.
7. Meixner, H., Mader, G., and Kleinschmidt, P. Infrared sensors based on the pyroelectric polymer polyvinylidene fluoride (PVDF). *Siemens Forsch. Entwicl. Ber. Bd.* 15(3), 105–114, 1986.
8. Fraden, J. Noncontact temperature measurements in medicine. In: *Bioinstrumentation and Biosensors*, D. Wise, ed. Marcel Dekker, New York, 1991. pp. 511–549.
9. Fraden, J. Infrared electronic thermometer and method for measuring temperature. U.S. patent 4,797,840, 1989.
10. Fraden, J. Motion detector, U.S. patent 4,769,545, 1988.
11. Astheimer, R.W. Thermistor infrared detectors. *Proc. SPIE* 443, 95–109, 1984.
12. Shie, J.-S. and Weng, P.K. Fabrication of micro-bolometer on silicon substrate by anizotropic etching technique. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*. IEEE, New York, 1991, pp. 627–630.
13. Vogl, T.P., Shifrin, G.A., and Leon, B.J. Generalized theory of metal-film bolometers. *J. Opt. Soc. Am.* 52, 957–964, 1962.

14. Fraden, J. Active far infrared detectors. In: *Temperature. Its Measurement and Control in Science and Industry*. J. F. Schooley, ed. American Institute of Physics, Washington, DC, 1992, Vol. 6, Part 2, pp. 831–836.
15. Fraden, J. Radiation thermometer and method for measuring temperature. U.S. patent 4,854,730, 1989.
16. Fraden, J. Active infrared motion detector and method for detecting movement. U.S. patent 4,896,039, 1990.
17. Mastrangelo, C.H. and Muller, R.S. Design and performance of constant-temperature circuits for microbridge-sensor applications. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*. IEEE, New York, 1991, pp. 471–474.

15

Radiation Detectors

Figure 3.41 of Chapter 3 shows a spectrum of the electromagnetic waves. On its left-hand side, there is a region of the γ-radiation. However, this is not the shortest possible length of the electromagnetic waves. In addition, a spontaneous radiation from the matter is not necessarily electromagnetic: There is the so-called nuclear radiation which is the emission of particles from the atomic nuclei. It can be of two types: the charged particles (α and β particles and protons) and uncharged particles, which are the neutrons. Some particles are complex like the α-particles, which are nuclei of helium atoms consisting of two neutrons; other particles are generally simpler, like the β-particles, which are either electrons or positrons. The γ- and X-rays belong to the nuclear type of electromagnetic radiation. In turn, X-rays depending on the wavelengths are divided into hard, soft, and ultrasoft rays. Ionizing radiations are given that name because as they pass through various media which absorb their energy, additional ions, photons, or free radicals are created.

Certain naturally occurring elements are not stable but slowly decompose by throwing away a portion of their nucleus. This is called *radioactivity*. It was discovered in 1896 by Henry Becquerel when he found that uranium atoms ($Z = 92$)[1] give off radiation which fogs photographic plates. In addition to the naturally occurring radioactivity, there are many man-made nuclei which are radioactive. These nuclei are produced in nuclear reactors, which may yield highly unstable elements. Regardless of the sources or ages of radioactive substances, they decay in accordance with the same mathematical law. The law is stated in terms of the number N of nuclei still undecayed and dN, the number of nuclei which decay in a small interval dt. It was proven experimentally that

$$dN = -\lambda N\, dt, \tag{15.1}$$

where λ is a decay constant specific for a given substance. From Eq. (15.1), it can be defined as the fraction of nuclei which decays in unit time:

$$\lambda = -\frac{1}{N}\frac{dN}{dt}. \tag{15.2}$$

[1] Z is the atomic number.

The SI unit of radioactivity is the *becquerel* (Bq) which is equal to the activity of a radionuclide decaying at the rate of one spontaneous transition per second. Thus, the becquerel is expressed in a unit of time: $\text{Bq} = \text{s}^{-1}$. To convert to the old historical unit, the *curie*, the becquerel should be multiplied by 3.7×10^{10} (Table A.4). The absorbed dose is measured in *grays* (Gy). A gray is the absorbed dose when the energy per unit mass imparted to matter by ionizing radiation is 1 joule per kilograms; that is, Gy = J/kg. When it is required to measure exposure to X- and γ-rays, the dose of ionizing radiation is expressed in coulombs per kilogram, which is an exposure resulting in the production of 1 C of electric charge per 1 kg of dry air. In SI, the unit C/kg replaces the older unit *roentgen*.

The function of any radiation detector depends on the manner in which the radiation interacts with the material of the detector itself. There are many excellent texts available on the subject of detecting radioactivity—for instance, Refs. [1] and [2].

There are three general types of radiation detector: the scintillation detector, the gaseous detector, and the semiconductor detector. Further, all detectors can be divided into two groups according to their functionality: the collision detector and the energy detector. The former merely detect the presence of a radioactive particle, whereas the latter can measure the radiative energy; that is, all detectors can be either quantitative or qualitative.

15.1 Scintillating Detectors

The operating principle of these detectors is based on the ability of certain materials to convert nuclear radiation into light. Thus, an optical photon detector in combination with a scintillating material can form a radiation detector. It should be noted, however, that despite the high efficiency of the conversion, the light intensity resulting from the radiation is extremely small. This demands photomultipliers to magnify signals to a detectable level.

The ideal scintillation material should possess the following properties:

1. It should convert the kinetic energy of charged particles into detectable light with a high efficiency.
2. The conversion should be linear; that is, the light produced should be proportional to the input energy over a wide dynamic range.
3. The postluminescence (the light decay time) should be short to allow fast detection.
4. The index of refraction of the material should be near that of glass to allow efficient optical coupling of the light to the photomultiplier tube.

The most widely used scintillators include the inorganic alkali halide crystals (of which sodium iodine is the favorite) and organic-based liquids and plastics. The inorganics are more sensitive, but generally slow, whereas organics are faster, but yield less light.

One of the major limitations of scintillation counters is their relatively poor energy resolution. The sequence of events which leads to the detection involves many

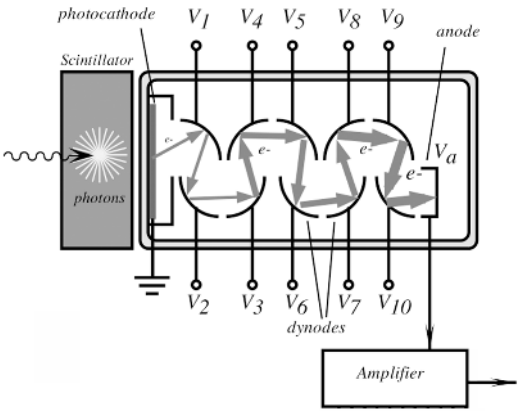


Fig. 15.1. Scintillation detector with a photomultiplier.

inefficient steps. Therefore, the energy required to produce one information carrier (a photoelectron) is on the order of 1000 eV or more, and the number of carriers created in a typical radiation interaction is usually no more than a few thousand. For example, the energy resolution for sodium iodine scintillators is limited to about 6% when detecting 0.662-MeV γ-rays and is largely determined by the photoelectron statistical fluctuations. The only way to reduce the statistical limit on energy resolution is to increase the number of information carriers per pulse. This can be accomplished by the use of semiconductor detectors, which are described in Section 15.2.4.

A general simplified arrangement of a scintillating sensor is shown in Fig. 15.1 in conjunction with a photomultiplier. The scintillator is attached to the front end of the photomultiplier (PM). The front end contains a photocathode which is maintained at a ground potential. There is a large number of special plates called *dynodes* positioned inside the PM tube in an alternating pattern, reminding one of the shape of a "venetian blind." Each dynode is attached to a positive voltage source in a manner that the farther the dynode from the photocathode, the higher is its positive potential. The last component in the tube is an anode, which has the highest positive potential, sometimes on the order of several thousand volts. All components of the PM are enveloped into a glass vacuum tube, which may contain some additional elements, like focusing electrodes, shields, and so forth.

Although the PM is called a photomultiplier, in reality it is an electron multiplier, as there are no photons, only electrons inside the PM tube during its operation. For the illustration, let us assume that a γ-ray particle has a kinetic energy of 0.5 MeV (megaelectron volt). It is deposited on the scintillating crystal resulting in a number of liberated photons. In thallium-activated sodium iodine, the scintillating efficiency is about 13%, therefore, a total of $0.5 \times 0.13 = 0.065$ MeV, or 65 keV, of energy is converted into visible light with an average energy of 4 eV. Therefore, about 15,000 scintillating photons are produced per gamma pulse. This number is too small to be detected by an ordinary photodetector; hence, a multiplication effect is required before the actual detection takes place. Of the 15,000 photons, probably about 10,000 reach the photocathode, whose quantum efficiency is about 20%. The photocathode serves to

convert incident light photons into low-energy electrons. Therefore, the photocathode produces about 2000 electrons per pulse. The PM tube is a linear device; that is, its gain is almost independent of the number of multiplied electrons.

Because all dynodes are at positive potentials (V_1 to V_{10}), an electron released from the photocathode is attracted to the first dynode, liberating several very low energy electrons at impact with its surface. Thus, a multiplication effect takes place at the dynode. These electrons will be easily guided by the electrostatic field from the first to the second dynode. They strike the second dynode and produce more electrons which travel to the third dynode, and so on. The process results in an increasing number of available electrons (avalanche effect). An overall multiplication ability of a PM tube is in the order of 10^6. As a result, about 2×10^9 electrons will be available at a high voltage anode (V_a) for the production of electric current. This is a very strong electric current which can be easily processed by an electronic circuit. A gain of a PM tube is defined as

$$G = \alpha \delta^N, \tag{15.3}$$

where N is the number of dynodes, α is the fraction of electrons collected by the PM tube, and δ is the efficiency of the dynode material (i.e., the number of electrons liberated at impact). Its value ranges from 5 to 55 for a high yield dynode. The gain is sensitive to the applied high voltage, because δ is almost a linear function of the interdynode voltage.

A new design of a photomultiplier is called the channel photomultiplier or CPM for short. It is the evolution of the classical photomultiplier tube. The modern CPM technology preserves the advantages of the classical PM while avoiding its disadvantages. Figure 15.2A shows the face plate with a photocathode, the bent channel amplification structure, and the anode. As in the PM of Fig. 15.1, photons in the CPM are converted inside the photocathode into photoelectrons and accelerated in a vacuum toward the anode by an electrical field. Instead of the complicated dynode structure, there is a bent, thin semiconductive channel which the electrons have to pass. Each time the electrons hit the wall of the channel, secondary electrons are

Fig. 15.2. Channel photomultiplier: cross-sectional view (A) and external view with potted encapsulation at left and without encapsulation at right (B). (Courtesy of Perkin-Elmer, Inc.)

emitted from the surface. At each collision, there is a multiplication of the secondary electrons resulting in an avalanche effect. Ultimately, an electron multiplication of 10^9 and more can be obtained. The resulting current can be read out at the anode. The CPM detector is potted with encapsulation material and is quite rugged compared to the fragile PM. Magnetic field disturbance is negligibly small. Figure 15.2B illustrates the CPM: on the left is a potted structure and on the right is the unpotted structure. An important advantage of the CPM technology is its very low background noise. The term "background noise" refers to the measured output signal in the absence of any incident light. With classical PMs, the background noise originating from the dynode structure is generally a non-negligible part of the total background. As a result the only effective source of background for the CPM is generated from the thermal emission of the photocathode. Because the CPM is manufactured in a monolithic semiconductive channel structure, no charge-up effects might occur as known from classical PMs with isolating glass bulbs. As a result, extremely stable background conditions are observed. No sudden bursts occur. Also, due to the absence of dynode noise, a very clean separation between an event created from a photoelectron and electronic noise can be performed. This leads into a high stability of the signal over time.

15.2 Ionization Detectors

These detectors rely on the ability of some gaseous and solid materials to produce ion pairs in response to the ionization radiation. Then, positive and negative ions can be separated in an electrostatic field and measured.

Ionization happens because upon passing at a high velocity through an atom, charged particles can produce sufficient electromagnetic forces, resulting in the separation of electrons, thus creating ions. Remarkably, the same particle can produce multiple ion pairs before its energy is expended. Uncharged particles (like neutrons) can produce ion pairs at collision with the nuclei.

15.2.1 Ionization Chambers

These radiation detectors are the oldest and most widely used. The ionizing particle causes ionization and excitation of gas molecules along its passing track. As a minimum, the particle must transfer an amount of energy equal to the ionization energy of the gas molecule to permit the ionization process to occur. In most gasses of interest for radiation detection, the ionization energy for the least tightly bound electron shells is between 10 and 20 eV [2]. However, there are other mechanisms by which the incident particle may lose energy within gas that do not create ions (e.g., moving gas electrons to a higher energy level without removing it). Therefore, the average energy lost by a particle per ion pair formed (called the *W value*) is always greater than the ionizing energy. The W value depends on the gas (Table 15.1), the type of radiation, and its energy.

In the presence of an electric field, the drift of the positive and negative charges represented by the ions end electrons constitutes an electric current. In a given volume

Table 15.1. W Values for Different Gases

	W Value (in eV/Ion Pair)	
Gas	Fast electrons	Alphas
A	27.0	25.9
He	32.5	31.7
N_2	35.8	36.0
Air	35.0	35.2
CH_4	30.2	29.0

Source: Ref. [2].

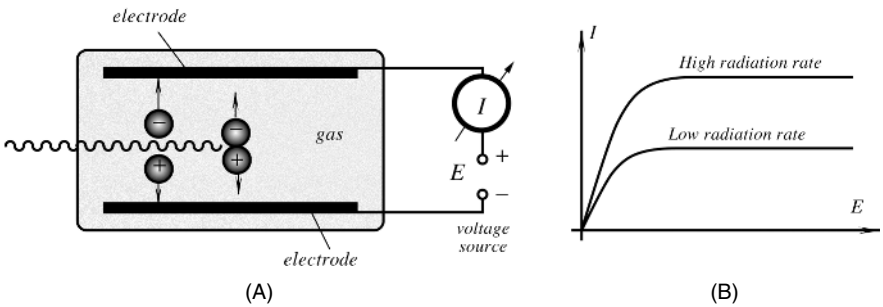


Fig. 15.3. Simplified schematic of an ionization chamber (A) and a current versus voltage characteristic (B).

of gas, the rate of the formation of the ion pair is constant. For any small volume of gas, the rate of formation will be exactly balanced by the rate at which ion pairs are lost from the volume, either through recombination or by diffusion or migration from the volume. If recombination is negligible and all charges are effectively collected, the steady-state current produced is an accurate measure of the rate of ion-pair formation. Figure 15.3 illustrates a basic structure of an ionizing chamber and the current versus voltage characteristic. A volume of gas is enclosed between the electrodes which produce an electric field. An electric current meter is attached in series with the voltage source E and the electrodes. There is no electrical conduction and no current under the no-ionization conditions. Incoming radiation produces, in the gas, positive and negative ions which are pulled by the electric field toward the corresponding electrodes, forming an electric current. The current versus voltage characteristic of the chamber is shown in Fig. 15.3B. At relatively low voltages, the ion recombination rate is strong and the output current is proportional to the applied voltage, because the higher voltage reduces the number of recombined ions. A sufficiently strong voltage completely suppress all recombinations by pulling all available ions toward the electrodes and the current becomes voltage independent. However, it still depends on the intensity of irradiation. This is the region called *saturation* and where the ionization chamber normally operates.

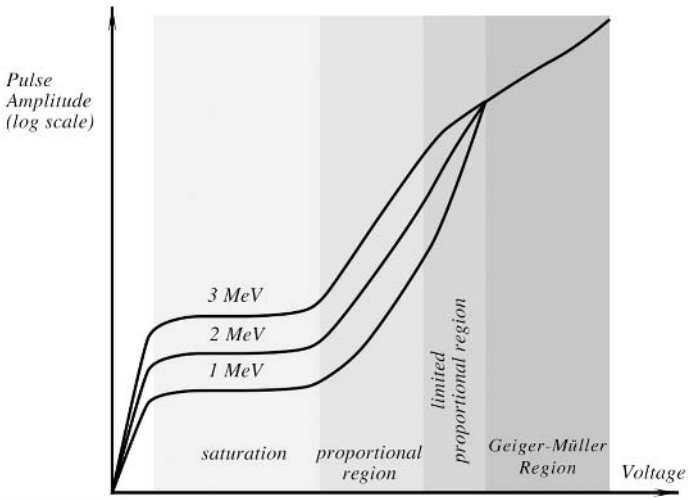


Fig. 15.4. Various operating voltages for gas-filled detectors. (Adapted from Ref. [2].)

15.2.2 Proportional Chambers

The proportional chamber is a type of a gas-filled detector which almost always operates in a pulse mode and relies on the phenomenon of gas multiplication. This is why these chambers are called the proportional counters. Due to gas multiplication, the output pulses are much stronger than in conventional ion chambers. These counters are generally employed in the detection and spectroscopy of low-energy X-radiation and for the detection of neutrons. Contrary to the ionization chambers, the proportional counters operate at higher electric fields which can greatly accelerate electrons liberated during the collision. If these electrons gain sufficient energy, they may ionize a neutral gas molecule, thus creating an additional ion pair. Hence, the process is of an avalanche type, resulting in a substantial increase in the electrode current. The name for this process is the Townsend avalanche. In the proportional counter, the avalanche process ends when the electron collides with the anode. Because in the proportional counter, the electron must reach the gas ionization level, there is a threshold voltage after which the avalanche process occurs. In typical gases at atmospheric pressure, the threshold field level is on the order of 10^6 V/m.

Differences between various gas counters are illustrated in Fig. 15.4. At very low voltages, the field is insufficient to prevent the recombination of ion pairs. In the saturation level, all ions drift to the electrodes. A further increase in voltage results in gas multiplication. Over some region of the electric field, the gas multiplication will be linear, and the collected charge will be proportional to the number of original ion pairs created during the ionization collision. An even further increase in the applied voltage can introduce nonlinear effects, which are related to the positive ions, due to their slow velocity.

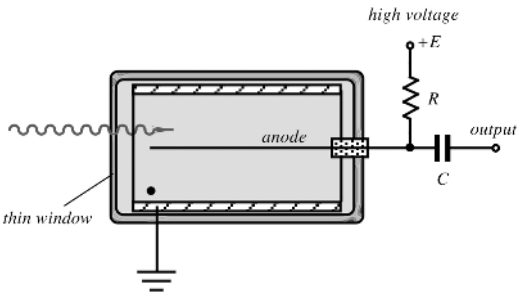


Fig. 15.5. Circuit of a Geiger–Müller counter. The symbol • indicates gas.

15.2.3 Geiger–Müller Counters

The Geiger–Müller(G-M) counter was invented in 1928 and is still in use because of its simplicity, low cost, and ease of operation. The G-M counter is different from other ion chambers by its much higher applied voltage (Fig. 15.4). In the region of the G-M operation, the output pulse amplitude does not depend on the energy of ionizing radiation and is strictly a function of the applied voltage. A G-M counter is usually fabricated in the form of a tube with an anode wire in the center (Fig. 15.5). The tube is filled with a noble gas, such as helium or argon. A secondary component is usually added to the gas for the purpose of *quenching*, which prevents the retriggering of the counter after the detection. The retriggering may cause multiple pulses instead of the desired one. The quenching can be accomplished by several methods, among which are a short-time reduction of the high voltage applied to the tube, use of high-impedance resistors in series with the anode, and the addition of the quench gas at concentrations of 5–10%. Many organic molecules possess the proper characteristics to serve as a quench gas. Of these, ethyl alcohol and ethyl formate have proven to be the most popular.

In a typical avalanche created by a single original electron, secondary ions are created. In addition to them, many excited gas molecules are formed. Within a few nanoseconds, these excited molecules return to their original state through the emission of energy in the form of ultraviolet (UV) photons. These photons play an important role in the chain reaction occurring in the G-M counter. When one of the UV photons interacts by photoelectric absorption in some other region of the gas, or at the cathode surface, a new electron is liberated which can subsequently migrate toward the anode and will trigger another avalanche. In a Geiger discharge, the rapid propagation of the chain reaction leads to many avalanches which initiate, at random, radial and axial positions throughout the tube. Secondary ions are therefore formed throughout the cylindrical multiplying region which surrounds the anode wire. Hence, the discharge grows to envelop the entire anode wire, regardless of the position at which the primary initiating event occurred.

Once the Geiger discharge reaches a certain level, however, collective effects of all individual avalanches come into play and ultimately terminate the chain reaction. This point depends on the number of avalanches and not on the energy of the initiating

particle. Thus, the G-M current pulse is always of the same amplitude, which makes the G-M counter just an indicator of irradiation, because all information on the ionizing energy is lost.

In the G-M counter, a single particle of a sufficient energy can create about 10^9–10^{10} ion pairs. Because a single ion pair formed within the gas of the G-M counter can trigger a full Geiger discharge, the counting efficiency for any charged particle that enters the tube is essentially 100%. However, the G-M counters are seldom used for counting neutrons because of a very low efficiency of counting. The efficiency of G-M counters for γ-rays is higher for those tubes constructed with a cathode wall of high-Z material. For instance, bismuth ($Z = 83$) cathodes have been widely used for the γ-detection in conjunction with gases of high atomic numbers, such as xenon and krypton, which yield a counting efficiency up to 100% for photon energies below about 10 keV.

15.2.4 Semiconductor Detectors

The best energy resolution in modern radiation detectors can be achieved in semiconductor materials, where a comparatively large number of carriers for a given incident radiation event occurs. In these materials, the basic information carriers are *electron–hole pairs* created along the path taken by the charged particle through the detector. The charged particle can be either primary radiation or a secondary particle. The electron–hole pairs in some respects are analogous to the ion pairs produced in the gas-filled detectors. When an external electric field is applied to the semiconductive material, the created carriers form a measurable electric current. The detectors operating on this principle are called a solid-state or semiconductor diode detectors. The operating principle of these radiation detectors is the same as that of the semiconductor light detectors. It is based on the transition of electrons from one energy level to another when they gain or lose energy. For the introduction to the energy-band structure in solids the reader should refer to Section 14.1 of Chapter 14.

When a charged particle passes through a semiconductor with the band structure shown in Fig. 14.1 of Chapter 14, the overall significant effect is the production of many electron–hole pairs along the track of the particle. The production process may be either direct or indirect, in that the particle produces high-energy electrons (or Δ rays) which subsequently lose their energy in production more electron–hole pairs. Regardless of the actual mechanism involved, what is of interest to our subject is that the average energy expended by the primary charged particle produces one electron–hole pair. This quantity is often called the "ionization energy." The major advantage of semiconductor detectors lies in the smallness of the ionization energy. Its value for silicon or germanium is about 3 eV, compared with 30 eV required to create an ion pair in typical gas-filled detectors. Thus, the number of charge carriers is about 10 times greater for the solid-state detectors for a given energy of a measured radiation.

To fabricate a solid-state detector, at least two contacts must be formed across a semiconductor material. For detection, the contacts are connected to the voltage source, which enables carrier movement. The use of a homogeneous Ge or Si, however, would be totally impractical. The reason for that is in an excessively high leakage

current caused by the material's relatively low resistivity (50 kΩ cm for silicon). When applied to the terminals of such a detector, the external voltage may cause a current which is three to five orders of magnitude greater than a minute radiation-induced electric current. Thus, the detectors are fabricated with the blocking junctions, which are reverse biased to dramatically reduce leakage current. In effect, the detector is a semiconductor diode which readily conducts (has low resistivity) when its anode (p side of a junction) is connected to a positive terminal of a voltage source and the cathode (an n side of the junction) to the negative. The diode conducts very little (it has very high resistivity) when the connection is reversed; thus, the name reverse biasing is implied. If the reverse bias is made very large (in excess of the manufacturer specified limit), the reverse leakage current abruptly increases (the breakdown effect), which often may lead to a catastrophic deterioration of detecting properties or to the device destruction.

Several configurations of silicon diodes are currently produced; among them are diffused junction diodes, surface-barrier diodes, ion-implanted detectors, epitaxial layer detectors, and others. The diffused junction and surface-barrier detectors find widespread applications for the detection of α-particles and other short-range radiation. A good solid-state radiation detector should possess the following properties:

1. Excellent charge transport
2. Linearity between the energy of the incident radiation and the number of electron–hole pairs
3. Absence of free charges (low leakage current)
4. Production of a maximum number of electron–hole pairs per unit of radiation
5. High detection efficiency
6. Fast response speed
7. Large collection area
8. Low cost

When using semiconductor detectors, several factors should be seriously considered. Among them are the dead-band layer of the detector and the possible radiation damage. If heavy charged particles or other weakly penetrating radiations enter the detector, there may be a significant energy loss before the particle reaches the active volume of the semiconductor. The energy can be lost in the metallic electrode and in a relatively thick silicon body immediately beneath the electrode. This thickness must be measured directly by the user if an accurate compensation is desirable. The simplest and most frequently used technique is to vary the angle of incidence of a monoenergetic charged particle radiation [2]. When the angle of incidence is zero (i.e., perpendicular to the detector's surface), the energy loss in the dead layer is given by

$$\Delta E_0 = \frac{dE_0}{dx} t, \tag{15.4}$$

where t is the thickness of the dead layer. The energy loss for an angle of incidence of Θ is

$$\Delta E(\theta) = \frac{\Delta E_0}{\cos\theta}. \tag{15.5}$$

Therefore, the difference between the measured pulse height for angles of incidence of zero and Θ is given by

$$E' = [E_0 - \Delta E_0] - [E_0 - \Delta E(\theta)] = \Delta E_0 \left(\frac{1}{\cos\theta} - 1 \right). \tag{15.6}$$

If a series of measurements is made as the angle of incidence is varied, a plot of E' as a function of $(1/\cos\Theta) - 1$ should be a straight line whose slope is equal to ΔE_0. Using tabular data for dE_0/dx for the incident radiation, the dead-layer thickness can be calculated from Eq. (15.4).

Any excessive use of the detectors may lead to some damage to the lattice of the crystalline structure, due to disruptive effects of the radiation being measured as it passes through the crystal. These effects tend to be relatively minor for lightly ionizing radiation (β-particles or γ-rays), but they can become quite significant under typical conditions of use for heavy particles. For example, prolonged exposure of silicon surface-barrier detectors to fusion fragments will lead to a measurable increase in leakage current and a significant loss in energy resolution of the detector. With extreme radiation damage, multiple peaks may appear in the pulse height spectrum recorded for monoenergetic particles.

As mentioned earlier, diffused junction diodes and surface-barrier diodes are not quite suitable for the detection of penetrating radiation. The major limitation is in the shallow active volume of these sensors, which rarely can exceed 2–3 mm. This is not nearly enough, for instance, for γ-ray spectroscopy. A practical method to make detectors for a more penetrating radiation is the so-called ion-drifting process. The approach consists of creating a thick region with a balanced number of donor impurities, which add either p or n properties to the material. Under ideal conditions, when the balance is perfect, the bulk material would resemble the pure (intrinsic) semiconductor without either properties. However, in reality, the perfect pn balance never can be achieved. In Si or Ge, the pure material with the highest possible purity tends to be of p type. To accomplish the desired compensation, the donor atoms must be added. The most practical compensation donor is lithium. The fabrication process involves a diffusing of lithium though the p crystal so that the lithium donors greatly outnumber the original acceptors, creating an n-type region near the exposed surface. Then, temperature is elevated and the junction is reverse biased. This results in a slow drifting of lithium donors into the p type for the near-perfect compensation of the original impurity. The process may take as long as several weeks. To preserve the achieved balance, the detector must be maintained at low temperature: 77K for the germanium detectors. Silicon has very low ion mobility; thus, the detector can be stored and operated at room temperature. However, the lower atomic number for silicon ($Z = 14$) as compared with germanium ($Z = 32$) means that the efficiency of silicon for the detection of γ-rays is very low and it is not widely used in general γ-ray spectroscopy.

A simplified schematic of a lithium-drifted detector is shown in Fig. 15.6A. It consists of three regions; the "intrinsic" crystal is in the middle. In order to create detectors of a larger active volume, the shape can be formed as a cylinder (Fig. 15.6B),

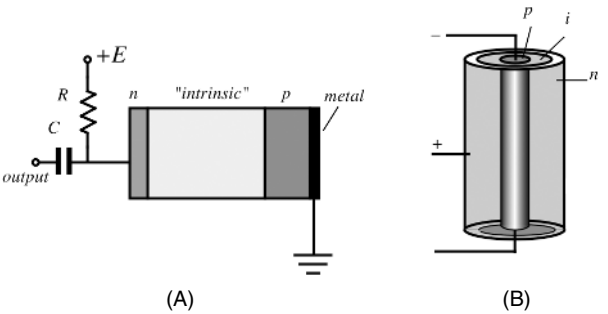


Fig. 15.6. Lithium-drifted PIN junction detector: (A) structure of the detector; (B) coaxial configuration of the detector.

Table 15.2. Detecting Properties of Some Semiconductive Materials

Material (Operating Temperature in K)	Z	Band Gap (eV)	Energy per Electron–Hole Pair, (eV)
Si (300)	14	1.12	3.61
Ge (77)	32	0.74	2.98
CdTe (300)	48–52	1.47	4.43
HgI_2 (300)	80–53	2.13	6.5
GaAs (300)	31–33	1.43	4.2

Source: Ref. [2].

where the active volumes of Ge up to 150 cm^3 can be realized. The germanium lithium-drifted detectors are designated Ge(Li).

Regardless of the widespread popularity of the silicon and germanium detectors, they are not the ideal from certain standpoints. For instance, germanium must always be operated at cryogenic temperatures to reduce thermally generated leakage current, and silicon is not efficient for the detection of γ-rays. There are some other semiconductors that are quite useful for detection of radiation at room temperatures. Among them are cadmium telluride (CdTe), mercuric iodine (HgI_2), gallium arsenide (GaAs), bismuth trisulfide (Bi_2S_3), and gallium selenide (GaSe). Useful radiation detector properties of some semiconductive materials are given in Table 15.2.

Probably the most popular at the time of this writing is cadmium telluride, which combines a relatively high Z-value (48 and 52) with a large enough band-gap energy (1.47 eV) to permit room-temperature operation. Crystals of high purity can be grown from CdTe to fabricate the intrinsic detector. Alternatively, chlorine doping is occasionally used to compensate for the excess of acceptors and to make the material a near-intrinsic type. Commercially available CdTe detectors range in size from 1 to 50 mm in diameter and can be routinely operated at temperatures up to 50°C without an excessive increase in noise. Thus, there are two types of CdTe detector available:

the pure intrinsic type and the doped type. The former has a high-volume resistivity up to $10^{10}\Omega$ cm, however, its energy resolution is not that high. The doped type has significantly better energy resolution; however, its lower resistivity ($10^{8}\Omega$ cm) leads to a higher leakage current. In addition, these detectors are prone to polarization, which may significantly degrade their performance.

In the solid-state detectors, it is also possible to achieve a multiplication effect as in the gas-filled detectors. An analog of a proportional detector is called an *avalanche detector*, which is useful for the monitoring of low-energy radiation. The gain of such a detector is usually in the range of several hundreds. It is achieved by creating high-level electric fields within a semiconductor. Also, the radiation PSDs are available whose operating principle is analogous to similar sensors functioning in the near-infrared region (see Section 7.5.6 of Chapter 7).

References

1. Evans, R.D. *The Atomic Nucleus*. McGraw-Hill, New York, 1955.
2. Knoll, G.F. *Radiation Detection and Measurement*. 3rd ed., John Wiley & Sons, New York, 1999.

This page intentionally left blank

16

Temperature Sensors

When a scientist thinks of something, he asks,–'Why?'
When an engineer thinks of something, he asks,–'Why not?'

Since prehistoric times people were aware of heat and tried to assess its intensity by measuring temperature. Perhaps the simplest and certainly the most widely used phenomenon for temperature sensing is thermal expansion. This forms the basis of the liquid-in-glass thermometers. For electrical transduction, different methods of sensing are employed. Among them are the resistive, thermoelectric, semiconductive, optical, acoustic, and piezoelectric detectors.

Taking a temperature essentially requires the transmission of a small portion of the object's thermal energy to the sensor, whose function is to convert that energy into an electrical signal. When a contact sensor (probe) is placed inside or on the object, heat conduction takes place through the interface between the object and the probe. The sensing element in the probe warms up or cools down; that is, it exchanges heat with the object. The same happens when heat is transferred by means of radiation: thermal energy in the form of infrared light is either absorbed by the sensor or liberated from it depending on the object's temperature and the optical coupling. Any sensor, no matter how small, will disturb the measurement site and thus cause some error in temperature measurement. This applies to any method of sensing: conductive, convective, and radiative. Thus, it is an engineering task to minimize the error by an appropriate sensor design and a correct measurement technique.

When a temperature sensor responds, two basic methods of the signal processing can be employed: *equilibrium* and *predictive*. In the equilibrium method, a temperature measurement is complete when no significant thermal gradient exists between the measured surface and the sensing element inside the probe. In other words, a thermal equilibrium is reached between the sensor and the object of measurement. In the predictive method, the equilibrium is not reached during the measurement time. It is determined beforehand, through the rate of the sensor's temperature change. After

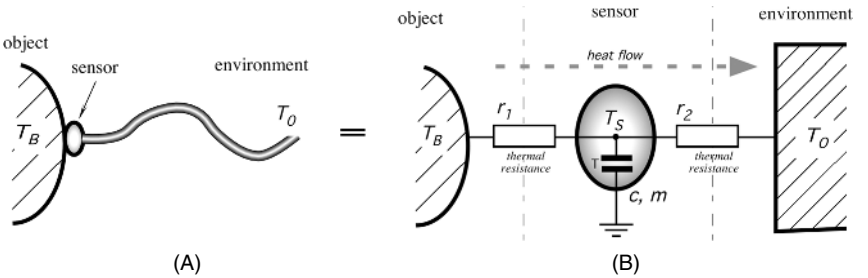


Fig. 16.1. A temperature sensor has thermal contacts with both the object and the connecting cable (A) and the equivalent thermal circuit (B).

the initial probe placement, reaching a thermal equilibrium between the object and the sensor may be a slow process, especially if the contact area is dry. Hence, the process of a temperature equalization may take significant time. For instance, a medical electronic thermometer may take temperature from a water bath within about 10 s, but it will be at least 3–5 minutes before temperature is measured axillary (under the armpit).

Let us discuss what affects the accuracy of a temperature measurement with a *contact* sensor. If a sensor is coupled not only to the object whose temperature it detects but also to some other items, an error is introduced. And to be sure, a temperature sensor is *always* attached to something else besides the object of measurement. An example of another item is a connecting cable (Fig. 16.1A). The sensor is coupled to the object (e.g., with an adhesive) and has its own temperature T_S. The object has temperature T_B. The goal of the equilibrium measurement is to bring T_S as close to T_B as possible. One end of the cable is connected to the sensor and the other end is subjected to ambient temperature T_0 which may be quite different from that of the object. The cable conducts both an electric signal and some portion of heat from or to the sensor. Figure 16.1B shows a thermal circuit that includes the object, sensor, environment, and thermal resistances r_1 and r_2. Thermal resistances should be clearly understood. They represent the ability of matter to conduct thermal energy and are inversely related to thermal conductivities; that is, $r = 1/\alpha$. If an object is warmer than the environment, heat flows in the direction indicated by the arrow.

The circuit in Fig. 16.1B resembles an electric circuit and indeed its properties can be computed by using the laws of electric circuits, such as Kirchhoff's[1] and Ohms laws. Note that a thermal capacitance is represented by a capacitor. Assuming that we wait sufficiently long and all temperatures are settled on some steady-state levels and also assuming that the object and environment temperatures are stable and not affected by their interconnection by the sensor, we may apply the law of conservation of energy. Consider that the thermal energy that flows from the object to the sensor is equal to the energy that outflows from the sensor to the environment. This allows us to write a balance equation:

$$\frac{T_B - T_S}{r_1} = \frac{T_B - T_0}{r_1 + r_2} \tag{16.1}$$

[1] Kirchoff's law was originally developed not for the electrical circuits but for plumbing.

from which we may find the sensor's temperature as

$$T_S = T_B - (T_B - T_0)\frac{r_1}{r_2} = T_B - \Delta T\frac{r_1}{r_2}, \tag{16.2}$$

where ΔT is a thermal gradient between the object and the surroundings. Let us take a closer look at Eq. (16.2). We can draw several conclusions from it. The first is that the sensor temperature T_S is always different from that of the object. The only exception is when the environment has the same temperature as the object (a special case when $\Delta T = T_B - T_0 = 0$). The second and the most important conclusion is that T_S will approach T_B at any temperature gradient ΔT when the ratio r_1/r_2 approaches zero. This means that for minimizing the measurement error, one must improve a thermal coupling between the object and the sensor and decouple the sensor from the surroundings as much as practical. Often, it is not easy to do.

In the above, we evaluated a static condition; now let us consider a dynamic case (i.e., when temperatures change with time). This occurs when either the object or the surrounding temperatures change or the sensor has just been attached to the object and its temperatures is not yet stabilized. When a temperature-sensing element comes in contact with the object, the incremental amount of transferred heat is proportional to a temperature gradient between that sensing element temperature T_S and that of the object T_B:

$$dQ = \alpha_1(T_B - T_S)\,dt, \tag{16.3}$$

where $\alpha_1 = 1/r_1$ is the thermal conductivity of the sensor–object interface. If the sensor has specific heat c and mass m, the absorbed heat is

$$dQ = mc\,dT. \tag{16.4}$$

If we ignore heat lost from the sensor to the environment through the connecting and supporting structure (assuming that $r_2 = \infty$), Eqs. (16.3) and (16.4) yield the first-order differential equation

$$\alpha_1(T_1 - T)\,dt = mc\,dT. \tag{16.5}$$

We define the thermal time constant τ_T as

$$\tau_T = \frac{mc}{\alpha_1} = mcr_1; \tag{16.6}$$

then, the differential equation takes the form

$$\frac{dT}{T_1 - T} = \frac{dt}{\tau_T}. \tag{16.7}$$

This equation has the solution

$$T_S = T_B - \Delta T e^{-t/\tau_T}, \tag{16.8}$$

where, initially, the sensor is assumed to be at temperature T_B. The time transient of the sensor's temperature, which corresponds to Eq. (16.8), is shown in Fig. 16.2A.

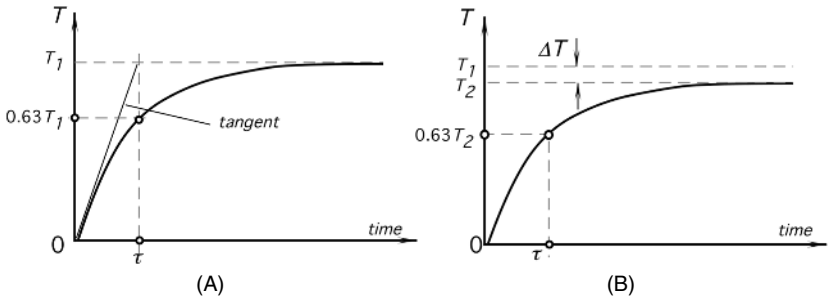


Fig. 16.2. Temperature changes of a sensing element: (A) the element is ideally coupled with the object (no heat loss); (B) the element has heat loss to its surroundings.

One time constant τ_T is equal to the time required for temperature T to reach 63.2% of the initial gradient ΔT. The smaller the time constant, the faster the sensor responds to a change in temperature.

If, in Eq. (16.8), $t \to \infty$, then the temperature of the sensor becomes equal to the temperature of the object: $T = T_1$. Theoretically, it takes an infinite amount of time to reach a perfect equilibrium between T_1 and T. However, because only a finite accuracy is usually required, for most practical cases a quasiequilibrium state may be considered after 5–10 time constants. For instance, after $t = 5\tau$, the sensor's temperature will differ from that of the object by 0.7% of the initial gradient ΔT_0, whereas after 10 time constants, it will be within 0.005%.

Now, if $r_2 \neq \infty$, the thermal time constant should be determined from

$$\tau_T = \frac{mc}{\alpha_1 + \alpha_2} = mc\frac{r_1}{1 + r_1/r_2} \tag{16.9}$$

and the sensor's response is shown in Fig. 16.2B. Note that the sensor's temperature never reaches exactly that of the object, no matter how long you wait.

A typical *contact* temperature sensor consists of the following components (Fig. 16.3A):

1. A sensing element: a material which is responsive to the change in its own temperature. A good element should have low specific heat, small mass, high thermal conductivity, and strong and predictable temperature sensitivity.

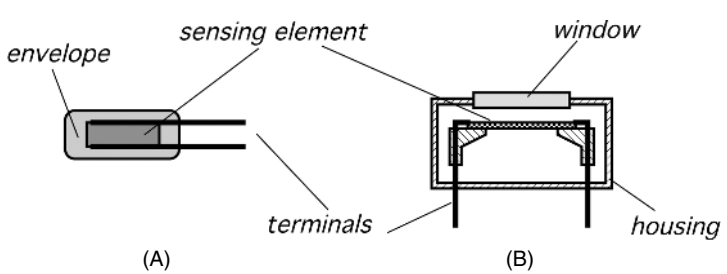


Fig. 16.3. General structures of temperature sensors: (A) contact sensor and (B) thermal radiation sensor (noncontact).

2. The contacts are the conductive pads or wires which interface between the sensing element and the external electronic circuit. The contacts should have the lowest possible thermal conductivity and electrical resistance. Also, they are often used to support the sensor.
3. A protective envelope is either a sheath or coating which physically separates a sensing element from the environment. A good envelope must have low thermal resistance (high thermal conductivity) and high electrical isolation properties. It must be impermeable to moisture and other factors which may spuriously affect the sensing element.

A *noncontact* temperature sensor (Fig. 16.3B) is an optical thermal radiation sensor whose designs are covered in detail in Chapter 14. Like a contact sensor, it also contains a sensing element which is responsive to its own temperature. The difference is in the method of a heat transfer from an object to the element: In a contact sensor, it is through thermal conduction via a physical contact, whereas in a noncontact sensor, it is through radiation or optically.

To improve the time response of a thermal radiation sensor, the thickness of the sensing element is minimized, whereas for better sensitivity, its surface area is maximized. In addition to a sensing element, the noncontact thermal sensor may have an optical window and a built-in interface circuit. The interior of the sensor's housing is usually filled with dry air or nitrogen.

All temperature sensors can be divided into two classes: the *absolute* sensors and the *relative* sensors. An absolute temperature sensor measures temperature which is referenced to the absolute zero or any other point on a temperature scale, such as 0°C (273.15° K), 25°C, and so forth. The examples of the absolute sensors are thermistors and resistance temperature detectors (RTDs). A relative sensor measures the temperature difference between two objects where one object is called a reference. An example of a relative sensor is a thermocouple.

16.1 Thermoresistive Sensors[2]

Sir Humphry Davy had noted as early as 1821 that electrical resistances of various metals depend on temperature [1]. Sir William Siemens, in 1871, first outlined the use of a platinum resistance thermometer. In 1887, Hugh Callendar published an article [2] in which he described how to practically use platinum temperature sensors. The advantages of thermoresistive sensors are in the simplicity of interface circuits, sensitivity, and long-term stability. All such sensors can be divided into three groups: RTDs, p-n junction detectors, and thermistors.

16.1.1 Resistance Temperature Detectors

This term is usually pertinent to metal sensors, fabricated either in the form of a wire or a thin film. The temperature dependence of resistivities of all metals and most alloys gives the opportunity to use them for temperature sensing (Table A.7). Although virtually all metals can be employed for sensing, platinum is used almost

[2] Also see Section 3.5.2 of Chapter 3.

Table 16.1. Temperature Reference Points

Point Description	°C
Triple point[a] of hydrogen	−259.34
Boiling point of normal hydrogen	−252.753
Triple point of oxygen	−218.789
Boiling point of nitrogen	−195.806
Triple point of argon	−189.352
Boiling point of oxygen	−182.962
Sublimation point of carbon dioxide	−78.476
Freezing point of mercury	−38.836
Triple point of water	0.01
Freezing point of water (water–ice mixture)	0.00
Boiling point of water	100.00
Triple point of benzoic acid	122.37
Freezing point of indium	156.634
Freezing point of tin	231.968
Freezing point of bismuth	271.442
Freezing point of cadmium	321.108
Freezing point of lead	327.502
Freezing point of zinc	419.58
Freezing point of antimony	630.755
Freezing point of aluminum	660.46
Freezing point of silver	961.93
Freezing point of gold	1064.43
Freezing point of copper	1084.88
Freezing point of nickel	1455
Freezing point of palladium	1554
Freezing point of platinum	1769

[a]The triple point is the equilibrium among the solid, liquid, and vapor phases.

exclusively because of its predictable response, long-term stability, and durability. Tungsten RTDs are usually applicable for temperatures over 600°C. All RTDs have positive temperature coefficients. Several types are available from various manufacturers:

1. Thin-film RTDs are often fabricated of a thin platinum or its alloys and deposited on a suitable substrate, such as a micromachined silicon membrane. The RTD is often made in a serpentine shape to ensure a sufficiently large length-to-width ratio.
2. Wire-wound RTDs, where the platinum winding is partially supported by a high-temperature glass adhesive inside a ceramic tube. This construction provides a detector with the most stability for industrial and scientific applications.

According to the International Practical Temperature Scale (IPTS-68), precision temperature instruments should be calibrated at reproducible equilibrium states of

Table 16.2. Temperature Differences Between IPTS-68 and ITS-90

t_{90}(°C)	−10	0	10	20	30	40
$T_{90} - t_{68}$(°C)	0.002	0.000	−0.002	−0.005	−0.007	−0.010

Source: Saunders, P. The International Temperature Scale of 1990, ITS-90. *WOCE Newsletter 10*, 1990.

some materials. This scale designated Kelvin temperatures by the symbol T_{68} and the Celsius scale by t_{68}. The International Committee for Weights and Measures adopted a new International Temperature Scale (ITS-90) during its meetings in September 1989. Its Celsius temperature designation is t_{90}.The difference between the two scales may be significant for some precision measurements (Table 16.2).

Equation (3.58) of Chapter 3 gives a best fit second-order approximation for platinum. In industry, it is customary to use separate approximations for the cold and hot temperatures. Callendar–van Dusen approximations represent the platinum transfer functions:

For the range from −200°C to 0°C,

$$R_t = R_0[1 + At + Bt^2 + Ct^3(t - 100)]. \tag{16.10}$$

For the range from 0°C to 630°C, it becomes identical to Eq. (3.58) of Chapter 3:

$$R_t = R_0(1 + At + Bt^2). \tag{16.11}$$

The constants A, B, and C are determined by the properties of platinum used in the construction of the sensor. Alternatively, the Callendar–van Dusen approximation can be written as

$$R_t = R_0 \left\{ 1 + \alpha \left[t - \delta \left(\frac{t}{100} \right) \left(\frac{t}{100} - 1 \right) - \beta \left(\frac{t}{100} \right)^3 \left(\frac{t}{100} - 1 \right) \right] \right\}, \tag{16.12}$$

where t is the temperature in °C and the coefficients are related to A, B, and C as

$$A = \alpha \left(1 + \frac{\delta}{100} \right), \qquad B = -\alpha\delta \times 10^{-4}, \qquad C = -\alpha\beta \times 10^{-8}. \tag{16.13}$$

The value of δ is obtained by calibration at a high temperature, [e.g., at the freezing point of zinc (419.58°C)] and β is obtained by calibration at a negative temperature.

To conform with ITS-90, the Callendar–van Dusen approximation must be corrected. The correction is rather complex and the user should refer for details to ITS-90. In different countries, some national specifications are applicable to RTDs. For instance, in Europe, these are the following: BS 1904: 1984; DIN 43760–1980; IEC 751: 1983. In Japan, it is JIS C1604-1981. In the United States, different companies have developed their own standards for α values. For example, SAMA Standard RC21-4-1966 specifies $\alpha = 0.003923°C^{-1}$, whereas in Europe, the DIN standard specifies $\alpha = 0.003850°C^{-1}$ and the British Aircraft industry standard is $\alpha = 0.003900°C^{-1}$.

Usually, RTDs are calibrated at standard points which can be reproduced in a laboratory with high accuracy (Table 16.1). Calibrating at these points allows for the precise determination of approximation constants α and δ.

Typical tolerances for the wire-wound RTDs is ± 10 mΩ, which corresponds to about ± 0.025°C. Giving high requirements to accuracy, packaging isolation of the device should be seriously considered. This is especially true at higher temperatures, at which the resistance of isolators may drop significantly. For instance, a 10-MΩ shunt resistor at 550°C results in a resistive error of about 3 mΩ, which corresponds to temperature error of -0.0075°C.

16.1.2 Silicon Resistive Sensors

Conductive properties of bulk silicon have been successfully implemented for the fabrication of temperature sensors with positive temperature coefficient (PTC) characteristics. Currently, silicon resistive sensors are often incorporated into the micromachined structures for temperature compensation or direct temperature measurement. There are also the discrete silicon sensors (e.g., the so-called KTY temperature detectors manufactured by Philips). These sensors have reasonably good linearity (which can be improved by the use of simple compensating circuits) and high long-term stability (typically, ± 0.05K per year). The PTC makes them inherently safe for operation in heating systems: A moderate overheating (below 200°C) results in RTD's resistance increase and self-protection.

Pure silicon, either polysilicon or single-crystal silicon, intrinsically has a negative temperature coefficient of resistance (NTC) (Fig. 18.1B of Chapter 18). However, when it is doped with an n-type impurity, in a certain temperature range its temperature coefficient becomes positive (Fig. 16.4). This is a result of the fall in the charge carrier mobility at lower temperatures. At higher temperatures, the number n of free charge carriers increases due to the number n_i of spontaneously generated charge carriers, and the intrinsic semiconductor properties of silicon predominate. Thus, at temperatures below 200°C, the resistivity ρ has a PTC; however, above 200°C, it becomes negative. The basic KTY sensor consists of an n-type silicon cell having approximate dimensions of $500 \times 500 \times 240$ µm, metallized on one side and having contact areas on the other side. This produces an effect of resistance "spreading," which causes a conical current distribution through the crystal, significantly reducing the sensor's dependence on manufacturing tolerances. A KTY sensor may be somewhat sensitive to current direction, especially at larger currents and higher temperatures. To alleviate this problem, a serially opposite design is employed where two of the sensors are connected with opposite polarities to form a dual sensor. These sensors are especially useful for automotive applications.

The typical sensitivity of a PTC silicon sensor is on the order of 0.7%/°C; that is, its resistance changes by 0.7% per every degree Celsius. As for any other sensor with a mild nonlinearity, the KTY sensor transfer function may be approximated by a second-order polynomial:

$$R_T = R_0[1 + A(T - T_0) + B(T - T_0)^2], \tag{16.14}$$

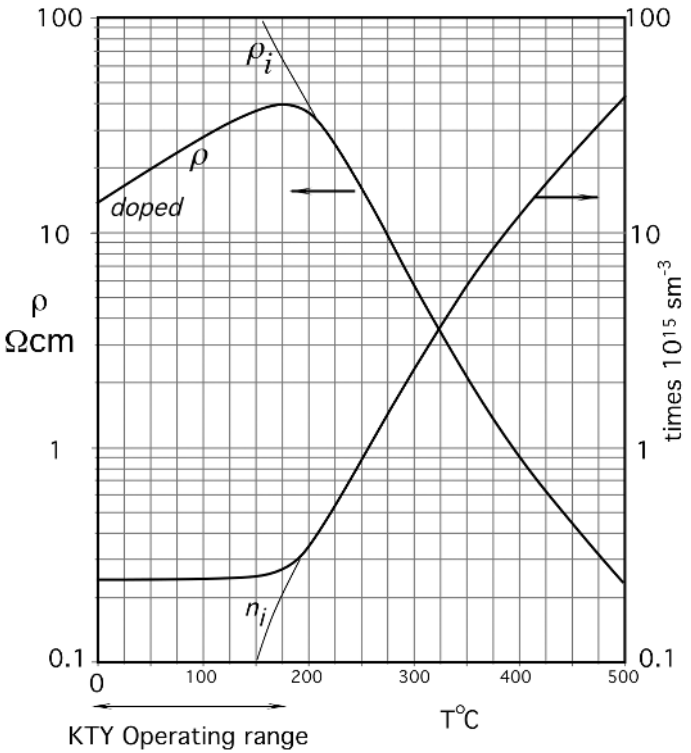


Fig. 16.4. Resistivity and number of free charge carriers for n-doped silicon.

where R_0 and T_0 are the resistance (Ω) and temperature (K) respectively, at a reference point. For instance, for the KTY-81 sensors operating in the range from −55°C to +150°C, the coefficients are $A = 0.007874\,\mathrm{K}^{-1}$ and $B = 1.874 \times 10^{-5}\,\mathrm{K}^{-2}$. A typical transfer function of the sensor is shown in Fig. 16.5.

16.1.3 Thermistors

The term *thermistor* is a contraction of the words *thermal* and *resistor*. The name is usually applied to metal-oxide sensors fabricated in the form of droplets, bars, cylinders, rectangular flakes, and thick films. A thermistor belongs to the class of absolute-temperature sensors; that is, it can measure temperature that is referenced to an absolute-temperature scale. All thermistors are divided into two groups: NTC (negative temperature coefficient) and PTC (positive temperature coefficient). Only the NTC thermistors are useful for precision temperature measurements.

16.1.3.1 NTC Thermistors

A conventional metal-oxide thermistor has a NTC; that is, its resistance decreases with the increase in temperature. The NTC thermistor's resistance, as of any resistor,

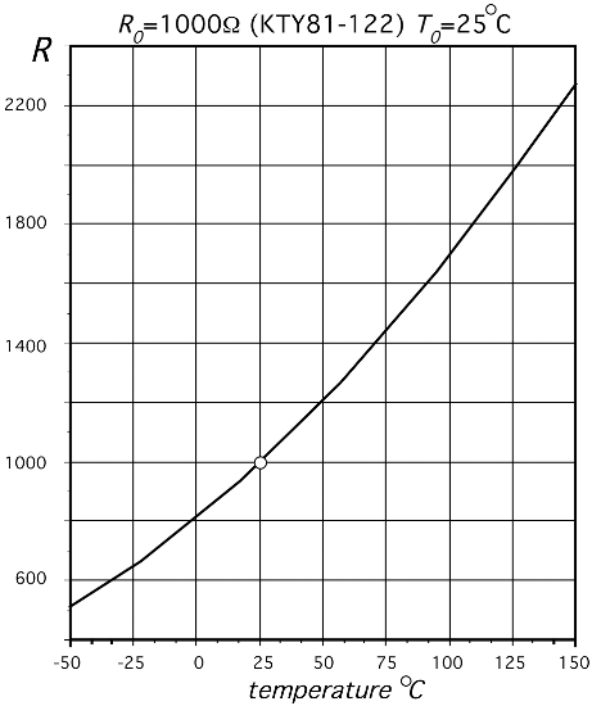


Fig. 16.5. Transfer function of a KTY silicon temperature sensor.

is determined by it physical dimensions and the material resistivity. The relationship between the resistance and temperature is highly nonlinear (Fig. 3.18 of Chapter 3).

Whenever high accuracy is required or the operating range is wide, the thermistor characteristics should not be taken directly from a manufacturer's data sheet. Typical tolerances of the nominal resistance (at 25°C) for mass-produced thermistors are rather wide: ±20% is quite common. Unless it was adjusted at the factory to a better tolerance, to reach a high accuracy each such thermistor needs to be individually calibrated over the operating temperature range. Manufacturers can trim a thermistor by grinding its body to a required dimension that directly controls the nominal value of resistance at a set temperature. This, however, increases cost. An alternative approach for an end user is to individually calibrate the thermistors. Calibration means that a thermistor has to be subjected to a precisely known temperature (a stirred water bath is often employed[3]) and its resistance is measured. This is repeated at several temperatures if a multipoint calibration is needed. Naturally, a thermistor calibration is as good as the accuracy of the reference thermometer used during the calibration. To measure the resistance of a thermistor, it is attached to a measurement circuit that passes through it electric current. Depending on the required accuracy and the

[3] Actually, water is not used. Mineral oil or Fluorinert® electronic fluid are more practical liquids.

production cost restrictions, a thermistor calibration can be based on the use of one of several known approximations (models) of its temperature response.

When a thermistor is used as an absolute-temperature sensor, we assume that all of its characteristics are based on the so-called "zero-power resistance", meaning that the electric current passing through a thermistor does not result in any temperature increase (self-heating) which may affect accuracy of measurement. A static temperature increase of a thermistor due to self-heating is governed by:

$$\Delta T_H = r\frac{N^2V^2}{S}, \tag{16.15}$$

where r is a thermal resistance to surroundings, V is the applied dc voltage during the resistance measurement, S is the resistance of a thermistor at a measured temperature, and N is a duty cycle of measurement (e.g., $N = 0.1$ means that constant voltage is applied to a thermistor only during 10% of the time). For a dc measurement, $N = 1$.

As follows from Eq. (16.15), a zero-power can be approached by selecting high-resistance thermistors, increasing the coupling to the object of measurement (reducing r), and measuring its resistance at low voltages applied during short time intervals. Later in this chapter, we will show the effects of self-heating on the thermistor response, but for now we assume that self-heating results in a negligibly small error.

To use a thermistor in the actual device, its transfer function (temperature dependence of a resistance) must be accurately established. Because that function is highly nonlinear and generally is specific for each particular device, an analytical equation connecting the resistance and temperature is highly desirable. Several mathematical models of a thermistor transfer function have been proposed. It should be remembered, however, that any model is only an approximation and, generally, the simpler the model, the lower the accuracy should be expected. On the other hand, for a more complex model, calibration and the use of a thermistor become more difficult. All present models are based on the experimentally established fact that the logarithm of a thermistor's resistance S relates to its absolute temperature T by a polynomial equation:

$$\ln S = A_0 + \frac{A_1}{T} + \frac{A_2}{T^2} + \frac{A_3}{T^3}, \tag{16.16}$$

From this basic equation, three computational models have been proposed.

16.1.3.1.1 Simple Model

Over a relatively narrow temperature range and assuming that some accuracy may be lost, we can eliminate two last terms in Eq. (16.16) and arrive at [3]

$$\ln S \cong A + \frac{\beta_m}{T}, \tag{16.17}$$

where A is a constant and β_m is another constant called the *material characteristic temperature* (in Kelvin). If a thermistor's resistance S_0 at a calibrating temperature T_0 is known, then the resistance–temperature relationship is expressed as:

$$S = S_0 e^{\beta_m(1/T - 1/T_0)}. \tag{16.18}$$

An obvious advantage of this model is a need to calibrate a thermistor at only one point (S_0 at T_0). However, this assumes that value of β_m is known beforehand, otherwise a two-point calibration is required to find the value of β_m:

$$\beta_m = \frac{\ln(S_1/S_0)}{(1/T_1 - 1/T_0)}, \tag{16.19}$$

where T_0 and S_0, and T_1 and S_1 are two pairs of the corresponding temperatures and resistances at two calibrating points on the curve described by Eq. (16.18). The value of β_m is considered temperature independent, but it may vary from part to part due to the manufacturing tolerances, which typically are within $\pm 1\%$. The temperature of a thermistor can be computed from its measured resistance S as

$$T = \left(\frac{1}{T_0} + \frac{\ln(S/S_0)}{\beta_m} \right)^{-1}. \tag{16.20}$$

The error of the approximation provided by Eq. (16.20) is small near the calibrating temperature, but it increases significantly with broadening of the operating range (Fig. 16.7).

β specifies a thermistor curvature, but it does not directly describe its sensitivity, which is a negative temperature coefficient, α. The coefficient can be found by differentiating Eq. (16.18):

$$\alpha_r = \frac{1}{S}\frac{dS}{dT} = -\frac{\beta}{T^2}. \tag{16.21}$$

It follows from Eq. (16.21) that the sensitivity depends on both β and temperature. A thermistor is much more sensitive at lower temperatures and its sensitivity drops quickly with a temperature increase. Equation (16.21) shows what fraction of a resistance S changes per degree of temperature. In the NTC thermistors, the sensitivity α varies over the temperature range from -2% (at the warmer side of the scale) to $-8\%/^\circ\text{C}$ (at the cooler side of the scale), which implies that an NTC thermistor is a very sensitive device, roughly an order of magnitude more temperature sensitive than a RTD. This is especially important for applications where a high output signal over a relatively narrow temperature range is desirable. An example is a medical electronic thermometer.

16.1.3.1.2 Fraden Model

In 1998, the author of this book proposed a further improvement of the simple model [4]. It is based on the experimental fact that the characteristic temperature β is not a constant but rather a function of temperature (Fig. 16.6). Depending on the manufacturer and type of thermistor, the function may have either a positive slope, as shown in Fig. 16.6, or a negative one. Ideally, β should not change with temperature, but that is just a special case that rarely happens in reality. When it does, the simple model provides a very accurate basis for temperature computation.

It follows from Eqs. (16.16) and (16.17) that the thermistor material characteristic temperature β can be approximated as

$$\beta = A_1 + BT + \frac{A_2}{T} + \frac{A_3}{T^2}, \tag{16.22}$$

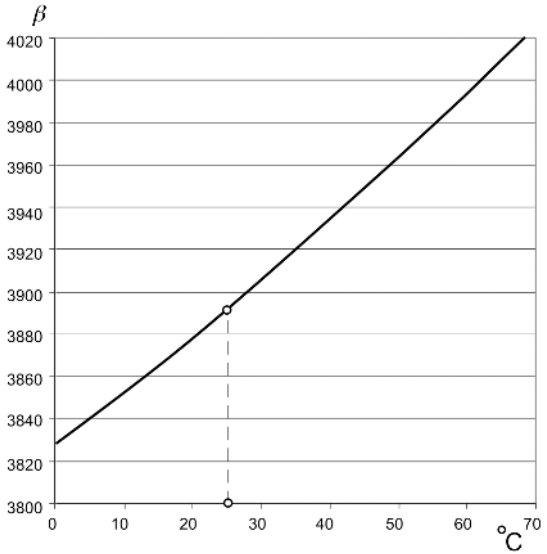


Fig. 16.6. Value of β changes with temperature.

where A and B are constants. The evaluation of this equation shows that the 3rd and 4th summands are very small as compared with the first two and for many practical cases can be removed. After elimination of two last terms, a model for the material constant can be represented as linear function of temperature:

$$\beta = A_1 + BT \tag{16.23}$$

Considering β as a linear function of temperature, the simple model can be refined to improve its fidelity. Because β is no longer a constant, for practical purposes its linear function should be defined through at least one fixed point at some temperature T_b and a slope γ. Then, Eq. (16.23) can be written as

$$\beta = \beta_b \left[1 + \gamma \left(T - T_b\right)\right], \tag{16.24}$$

where β_b is attributed to the temperature T_b. The coefficient γ is the normalized change (a slope) in β per degree Celsius:

$$\gamma = \left(\frac{\beta_x}{\beta_y} - 1\right) \frac{1}{T_c - T_a}, \tag{16.25}$$

where β_x and β_y are two material characteristic temperatures at two T_a and T_c characterizing temperatures[4] To determine γ, three characterizing temperature points are required (T_a , T_b and T_c), however, the value of γ does not need to be characterized for each individual thermistor. The value of γ depends on the thermistor material

[4] Note that β and T are in Kelvin. When temperature is indicated as t, the scale is in Celsius.

and the manufacturing process, so it may be considered more or less constant for a production lot of a particular type of a thermistor. Thus, it is usually sufficient to find γ for a production lot or type of a thermistor rather than for each individual sensor.

By substituting Eq. (16.23) into Eq. (16.16), we arrive at a model of a thermistor:

$$\ln S \cong A + \frac{\beta_m \left[1 - \gamma \left(T_b - T\right)\right]}{T}. \tag{16.26}$$

Solving Eq. (16.26) for resistance S, we obtain the equation representing the thermistor's resistance as a function of its temperature:

$$S = S_0 e^{\beta_m [1+\gamma (T-T_0)](1/T-1/T_0)}, \tag{16.27}$$

where S_0 is the resistance at calibrating temperature T_0 and β_m is the characteristic temperature defined at two calibrating temperatures T_0 and T_1 [Eq. 16.19)]. This is similar to a simple model of Eq. (16.18) with an introduction of an additional constant γ. Even though this model requires three points to define γ for a production lot, each individual thermistor needs to be calibrated at two points. This makes the Fraden model quite attractive for low-cost, high-volume applications which, at the same time, require higher accuracy. Note that the calibrating temperatures T_0 and T_1 preferably should be selected closer to the ends of the operating range and for the characterization, temperature T_B should be selected near the middle of the operating range. See Table 16.3 for the practical equations for using this model.

16.1.3.1.3 The Steinhart and Hart Model

Steinhart and Hart in 1968 proposed a model for the oceanographic range from −3°C to 30°C [5] which, in fact, is useful for a much broader range. The model is based on Eq. (16.16), from which temperature can be calculated as

$$T = \left[\alpha_0 + \alpha_1 \ln S + \alpha_2 \left(\ln S\right)^2 + \alpha_3 \left(\ln S\right)^3\right]^{-1}. \tag{16.28}$$

Steinhart and Hart showed that the square term can be dropped without any noticeable loss in accuracy; thus, the final equation becomes

$$T = \left[b_0 + b_1 \ln S + b_3 \left(\ln S\right)^3\right]^{-1}. \tag{16.29}$$

The correct use of Eq. (16.29) assures accuracy in a millidegree range from 0°C to 70°C [6]. To find coefficients b for the equation, a system of three equations should be solved after the thermistor is calibrated at three temperatures (Table 16.3). Because of the very close approximation, the Steinhart and Hart model became an industry standard for calibrating precision thermistors. Extensive investigation of its accuracy has demonstrated that even over a broad temperature range, the approximation error does not exceed the measurement uncertainty of a couple of millidegrees [7]. Nevertheless, a practical implementation of the approximation for the mass produced instruments is significantly limited by the need to calibrate each sensor at three or more temperature points.

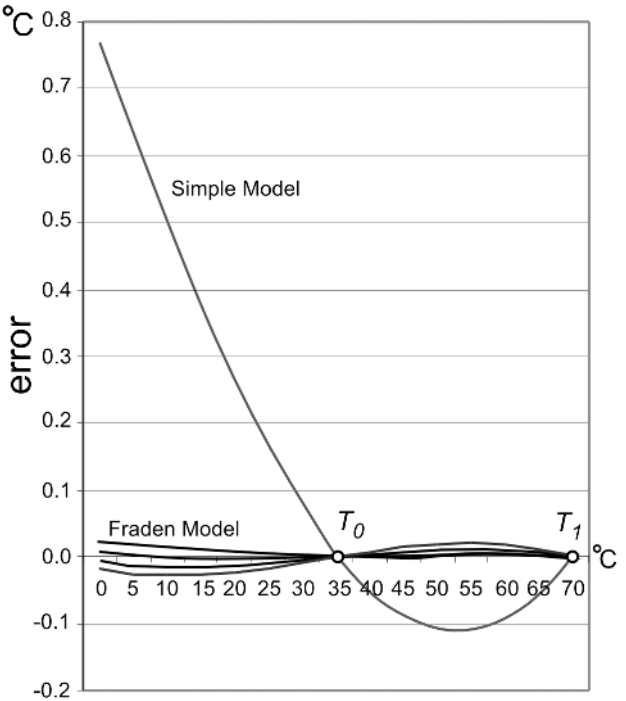


Fig. 16.7. Errors of a simple model and Fraden model for four thermistors calibrated at two temperature points (t_0 and t_1) to determine β_m. Errors of a Steinhart–Hart model are too small to be shown on this scale.

A practical selection of the appropriate model depends on the required accuracy and cost constraints. The cost is affected by the number of points at which the sensor must be calibrated. A calibration is time-consuming and thus expensive. A complexity of mathematical computations is not a big deal thanks to the computational power of modern microprocessors. When the accuracy demand is not high, or cost is of a prime concern, or the application temperature range is narrow (typically ±5–10°C from the calibrating temperature), the simple model is sufficient. The Fraden model is preferred when low cost and higher accuracy is a must. The Steinhart–Hart model should be used when the highest possible accuracy is required but the cost is not a major limiting factor (Fig. 16.7).

To use the simple model, you need to know the values of β_m and the thermistor resistance S_0 at a calibrating temperature T_0. To use the Fraden model, you need to know the value of γ also, which is not unique for each thermistor but is unique for a lot or a type. For the Steinhart–Hart model, you need to know three resistances at three calibrating temperatures. Table 16.3 provides the equations for calibrating and computing temperatures from the thermistor resistances. For all three models, a series of computations is required if the equations to be resolved directly. However, in most practical cases, these equations can be substituted by look-up tables. To minimize the size of a look-up table, a piecewise linear approximation can be employed.

Table 16.3. NTC Thermistor. Practical Use of Three Models[a]

	Simple Model	Fraden Model	Steinhart–Hart Model
Maximum Error from 0°C to 70°C	±0.7°C	±0.03°C	±0.003°C
Number of characterizing temperatures	2	3	0
Number of calibrating temperatures	2	2	3
Resistance–temperature dependence	$S = S_0 e^{\beta_m (1/T - 1/T_0)}$	$S = S_0 e^{\beta_0 [1+\gamma(T-T_0)](1/T - 1/T_0)}$	$S = e^{(A_0 + A_1/T + A_2/T^2 + A_3/T^3)}$
Characterizing a Production Lot or Type of Thermistors			
Characterizing points	No characterization required for a two-point calibration	S_a at T_a, S_b at T_b, and S_c at T_c for a temperature range from T_a to T_c where T_b is in the middle of the range	No characterization required
Characterizing factor		$\gamma = \left(\frac{\beta_x}{\beta_y} - 1\right) \frac{1}{T_c - T_a}$, where $\beta_x = \frac{\ln(S_c/S_b)}{(1/T_c - 1/T_b)}$, $\beta_y = \frac{\ln(S_a/S_b)}{(1/T_a - 1/T_b)}$	
Calibrating an Individual Thermistor			
Calibrating points	S_0 at T_0 and S_1 at T_1	S_0 at T_0 and S_1 at T_1	S_1 at T_1, S_2 at T_2, and S_3 at T_3
Analytic Computation of Temperature T (in Kelvin) from Resistance S			
Insert resistance S, the characterizing factors and the calibrating factors into the equation	$T = \left(\frac{1}{T_0} + \frac{\ln(S_1/S_0)}{\beta_m}\right)^{-1}$ where $\beta_m = \frac{\ln(S_1/S_0)}{(1/T_1 - 1/T_0)}$	$T = \left(\frac{1}{T_0} + \frac{\ln(S/S_0)}{\beta_m[1-\gamma(T_1 - T_r)]}\right)^{-1}$ where $T_r = \left(\frac{1}{T_0} + \frac{\ln(S/S_0)}{\beta_m}\right)^{-1}$ and $\beta_m = \frac{\ln(S_1/S_0)}{(1/T_1 - 1/T_0)}$	$T = [A + B \ln S + C(\ln S)^3]^{-1}$ where $C = \left(G - \frac{ZH}{F}\right)\left[\left(\ln S_1^3 - \ln S_2^3\right) - \frac{Z}{F}\left(\ln S_1^3 - \ln S_3^3\right)\right]^{-1}$ $B = Z^{-1}\left[G - C\left(\ln S_1^3 - \ln S_2^3\right)\right]$ $A = T_1^{-1} - C \ln S_1^3 - B \ln S_1$ $Z = \ln S_1 - \ln S_2$, $F = \ln S_1 - \ln S_3$ $H = T_1^{-1} - T_3^{-1}$, $G = T_1^{-1} - T_2^{-1}$

[a]A thermistor type or lot should be characterized first to find the *characterizing* factors. An individual thermistor is calibrated to determine the *calibrating factors*. To compute any temperature T, measure the thermistor resistance S and calculate the temperature with use of the characterizing and calibrating factors. All temperatures are in Kelvin.

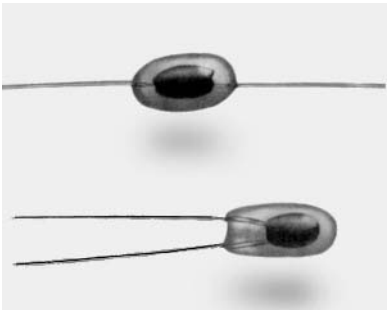


Fig. 16.8. Glass-coated radial and axial bead thermistors.

16.1.3.1.4 Fabrication of NTC Thermistors

Generally, the NTC thermistors can be classified into three major groups depending on the method by which they are fabricated. The first group consists of bead-type thermistors. The beads may be bare or coated with glass epoxy (Fig. 16.8), or encapsulated into a metal jacket. All of these beads have platinum alloy lead wires which are sintered into the ceramic body. When fabricated, a small portion of a mixed metal oxide with a suitable binder is placed onto parallel leadwires, which are under slight tension. After the mixture has been allowed to dry or has been partly sintered, the strand of beads is removed from the supporting fixture and placed into a tubular furnace for the final sintering. The metal oxide shrinks onto the lead wires during this firing process and forms an intimate electrical bond. Then, the beads are individually cut from the strand and are given an appropriate coating.

Another type of thermistor is a chip thermistor with surface contacts for the lead wires. Usually, the chips are fabricated by a tape-casting process, with subsequent screenprinting, spraying, painting, or vacuum metallization of the surface electrodes. The chips are either bladed or cut into the desired geometry. If desirable, the chips can be ground to meet the required tolerances.

The third type of thermistor is fabricated by the depositing semiconductive materials on a suitable substrate, such as glass, alumina, silicon, and so forth. These thermistors are preferable for integrated sensors and for a special class of thermal infrared detectors.

Among the metallized surface contact thermistors, flakes and uncoated chips are the least stable. A moderate stability may be obtained by epoxy coating. The bead type with lead wires sintered into the ceramic body permits operation at higher temperatures, up to 550°C. The metallized surface contact thermistors usually are rated up to 150°C. Whenever a fast response time is required, bead thermistors are preferable; however, they are more expensive than the chip type. Also, the bead thermistors are more difficult to trim to a desired nominal value. Trimming is usually performed by mechanical grinding of a thermistor at a selected temperature (usually 25°C) to change its geometry and thus to bring its resistance to a specified value.

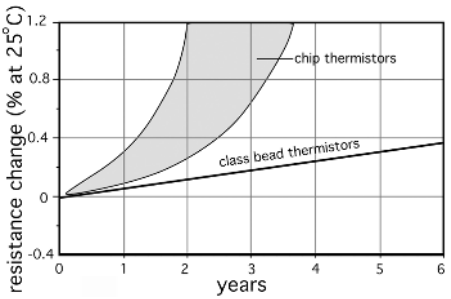


Fig. 16.9. Long-term stability of thermistors.

While using the NTC thermistors, one must not overlook possible sources of error. One of them is aging, which for the low-quality sensors may be as large as +1%/year. Figure 16.9 shows typical percentage changes in resistance values for the epoxy-encapsulated chip thermistors as compared with the sintered-glass-encapsulated thermistors. A good environmental protection and preaging is a powerful method of sensor characteristic stabilizing. During preaging, the thermistor is maintained at +300°C for at least 700 h. For better protection, it may be further encapsulated in a stainless-steel jacket and potted with epoxy.

16.1.3.2 Self-Heating Effect in NTC Thermistors

As was mentioned earlier, for NTC thermistor performance, a self-heating effect should not be overlooked. A thermistor is an active type of a sensor; that is, it does require an excitation signal for its operation. The signal is usually either dc or ac passing through the thermistor. The electric current causes a Joule heating and a subsequent increase in temperature. In many applications, this is a source of error which may result in an erroneous determination of the temperature of a measured object. In some applications, the self-heating is successfully employed for sensing fluid flow, thermal radiation, and other stimuli.

Let us now analyze the thermal events in a thermistor, when electric power is applied. Figure 16.10A shows a voltage source E connected to a thermistor R_T thorough a current-limiting resistor R. When electric power P is applied to the circuit

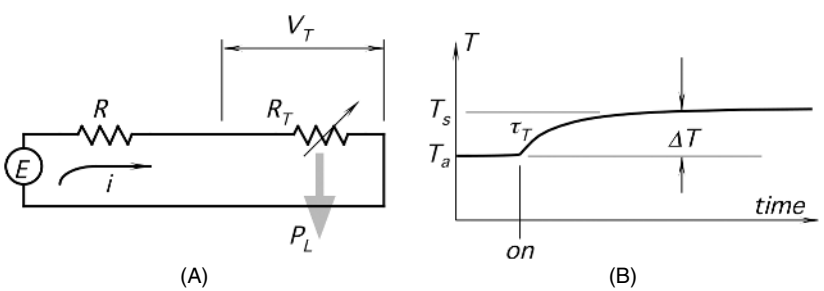


Fig. 16.10. (A) Current through thermistor causes self-heating; (B) temperature of thermistor rises with thermal time constant τ_T. P_L is the thermal power lost to the surroundings.

(moment *on* in Fig. 16.10B), the rate at which energy is supplied to the thermistor must be equal the rate at which energy H_L is lost plus the rate at which energy H_S is absorbed by the thermistor body. The absorbed energy is stored in the thermistor's thermal capacity C. The power balance equation is

$$\frac{dH}{dt} = \frac{dH_L}{dt} + \frac{dH_S}{dt}. \tag{16.30}$$

According to the law of conservation of energy, the rate at which thermal energy is supplied to the thermistor is equal to the electric power delivered by the voltage source E:

$$\frac{dH}{dt} = P = \frac{V_{T^2}}{R} = V_T i, \tag{16.31}$$

where V_T is the voltage drop across the thermistor.

The rate at which thermal energy is lost from the thermistor to its surroundings is proportional to the temperature gradient ΔT between the thermistor and surrounding temperature T_a:

$$P_L = \frac{dH_L}{dt} = \delta \Delta T = \delta (T_S - T_a), \tag{16.32}$$

where $\delta$is the so-called *dissipation factor* which is equivalent to a thermal conductivity from the thermistor to its surroundings. It is defined as the ratio of dissipated power and a temperature gradient (at a given surrounding temperature). The factor depends on the sensor design, length and thickness of lead wires, thermistor material, supporting components, thermal radiation from the thermistor surface, and relative motion of medium in which the thermistor is located.

The rate of heat absorption is proportional to thermal capacity of the sensor assembly:

$$\frac{dH_S}{dt} = C\frac{dT_S}{dt}. \tag{16.33}$$

This rate causes the thermistor's temperature T_S to rise above its surroundings. Substituting Eqs. (16.32) and (16.33) into Eq. (16.31), we arrive at

$$\frac{dH}{dt} = P = Ei = \delta (T_S - T_a) + C\frac{dT_S}{dt}. \tag{16.34}$$

This is a differential equation describing the thermal behavior of the thermistor. Let us now solve it for two conditions. The first condition is the constant electric power supplied to the sensor: $P = \text{const}$. Then, the solution of Eq. (16.34) is

$$\Delta T = (T_S - T_a) = \frac{P}{\delta}\left[1 - e^{-\delta/Ct}\right], \tag{16.35}$$

where e is the base of natural logarithms. This solution indicates that upon applying electric power, the temperature of the sensor will rise exponentially above ambient. This specifies a transient condition which is characterized by a thermal time constant $\tau_T = C(1/\delta)$. Here, the value of $1/\delta = r_T$ is the thermal resistance between the sensor and its surroundings. The exponential transient is shown in Fig. 16.10B.

Upon waiting sufficiently long to reach a steady-state level T_S, the rate of change in Eq. (16.34) becomes equal to zero ($dT_S/dt = 0$); then, the rate of heat loss is equal to supplied power:

$$\delta(T_S - T_a) = \delta \Delta T = V_T i. \tag{16.36}$$

If by selecting a low supply voltage and high resistances, the current i is made very low, the temperature rise ΔT can be made negligibly small, and self-heating is virtually eliminated. Then, from Eq. (16.34),

$$\frac{dT_S}{dt} = -\frac{\delta}{C}(T_S - T_a). \tag{16.37}$$

The solution of this differential equation yields an exponential function [Eq. (16.8)], which means that the sensor responds to the change in environmental temperature with time constant τ_T. Because the time constant depends on the sensor's coupling to the surroundings, it is usually specified for certain conditions; for instance, $\tau_T = 1$ s at 25°C in still air or 0.1 s at 25°C in stirred water. It should be kept in mind that the above analysis represents a simplified model of the heat flows. In reality, a thermistor response has a somewhat nonexponential shape.

All thermistor applications require the use of one of three basic characteristics:

1. The resistance versus temperature characteristic of the NTC thermistor is shown in Fig. 16.12. In most of the applications based on this characteristic, the self-heating effect is undesirable. Thus, the nominal resistance R_{T_0} of the thermistor should be selected high and its coupling to the object should be maximized (increase in δ). The characteristic is primarily used for sensing and measuring temperature. Typical applications are contact electronic thermometers, thermostats, and thermal breakers.
2. The current versus time (or resistance versus time) as shown in Fig. 16.10B.
3. The voltage versus current characteristic is important for applications where the self-heating effect is employed, or otherwise cannot be neglected. The power-supply-loss balance is governed by Eq. (16.36). If variations in δ are small (which is often the case) and the resistance versus temperature characteristic is known, then Eq. (16.36) can be solved for the static voltage versus current characteristic. That characteristic is usually plotted on log-log coordinates, where lines of constant resistance have a slope of $+1$ and lines of constant power have slope of -1 (Fig. 16.11).

At very low currents (left side of Fig. 16.11), the power dissipated by the thermistor is negligibly small and the characteristic is tangential to a line of constant resistance of the thermistor at a specified temperature. Thus, the thermistor behaves as a simple resistor; that is, the voltage drop V_T is proportional to current i.

As the current increases, the self-heating increases as well. This results in a decrease in the resistance of the thermistor. Because the resistance of the thermistor is no longer constant, the characteristics start to depart from the straight line. The slope of the characteristic (dV_T/di), which is the resistance, drops with the increase in current. The current increase leads to a further resistance drop which, in turn,

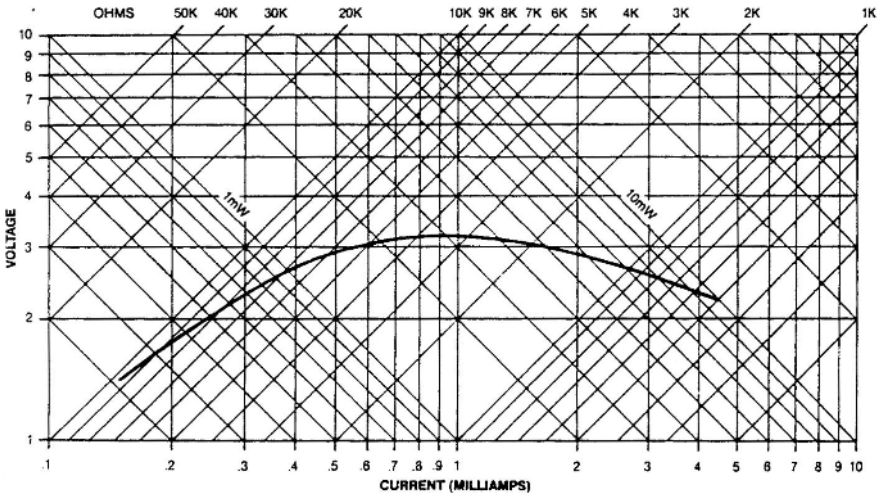


Fig. 16.11. Voltage–current characteristic of an NTC thermistor in still air at 25°C; the curvature of the characteristic is due to the self-heating effect.

increases the current. Eventually, the current will reach its maximum value i_p at a voltage maximum value V_p. It should be noted that at this point, a resistance of the thermistor is zero. A further increase in current i_p will result in a continuing decrease in the slope, which means that the resistance has a negative value (right side of Fig. 16.11). An even further increase in current will produce another reduction of resistance, where lead wire resistance becomes a contributing factor. A thermistor should never be operated under such conditions. A thermistor manufacturer usually specifies the maximum power rating for thermistors.

According to Eq. (16.36), self-heating thermistors can be used to measure variations in δ, ΔT, or V_T. The applications where δ varies include vacuum manometers (Pirani gauges), anemometers, flow meters, fluid level sensors, and so forth. Applications where ΔT is the stimulus include microwave power meters, AFIR detectors, and so forth. The applications where V_T varies are in some electronic circuits: automatic gain control, voltage regulation, volume limiting, and so forth.

16.1.3.3 PTC Thermistors

All metals may be called PTC materials, however, their temperature coefficients of resistivity (TCR) are quite low (Table A.7). An RTD as described earlier also has a small PTC. In contrast, ceramic PTC materials in a certain temperature range are characterized by a very large temperature dependence. The PTC thermistors are fabricated of polycrystalline ceramic substances, where the base compounds, usually barium titanate or solid solutions of barium and strontium titanate (highly resistive materials), are made semiconductive by the addition of dopants [8]. Above the Curie temperature of a composite material, the ferroelectric properties change rapidly, re-

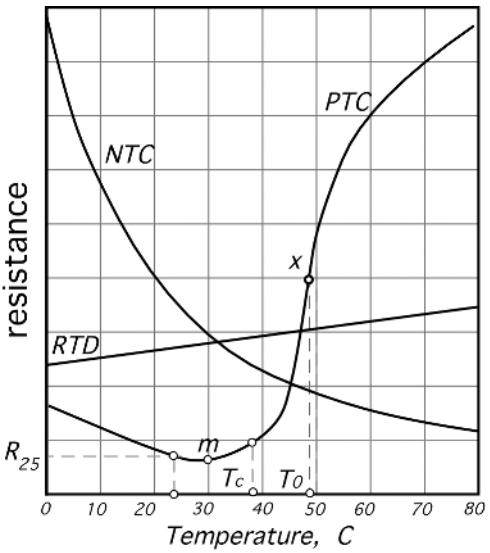


Fig. 16.12. Transfer functions of PTC and NTC thermistors as compared with an RTD.

sulting in a rise in resistance, often several orders of magnitude. A typical transfer function curve for the PTC thermistor is shown in Fig. 16.12 in a comparison with the NTC and RTD responses. The shape of the curve does not lend itself to an easy mathematical approximation; therefore, manufacturers usually specify PTC thermistors by a set of numbers:

1. Zero power resistance, R_{25}, at 25°C, where self-heating is negligibly small.
2. Minimum resistance R_m is the value on the curve where the thermistor changes its TCR from positive to negative value (point m).
3. Transition temperature T_τ is the temperature where resistance begins to change rapidly. It coincides approximately with the Curie point of the material. A typical range for the transition temperatures is from −30°C to +160°C (Keystone Carbon Co.).
4. TCR is defined in a standard form:

$$\alpha = \frac{1}{R}\frac{\Delta R}{\Delta T}. \tag{16.38}$$

   The coefficient changes very significantly with temperature and often is specified at point x (i.e., at its highest value), which may be as large as 2/°C (meaning the change in resistance is 200% per °C).
5. Maximum voltage $E_{\max}$ is the highest value the thermistor can withstand at any temperature.
6. Thermal characteristics are specified by a thermal capacity, a dissipation constant δ (specified under given conditions of coupling to the environment), and a thermal time constant (defines speed response under specified conditions).

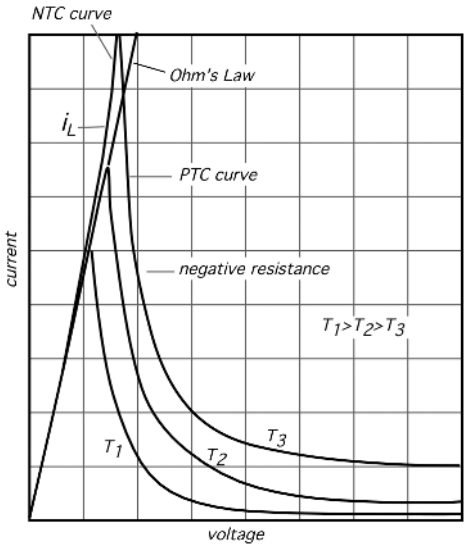


Fig. 16.13. Volt–ampere characteristic of a PTC thermistor.

It is important to understand that for the PTC thermistors, two factors play a key role: environmental temperature and a self-heating effect. Either one of these two factors shifts the thermistor's operating point.

The temperature sensitivity of the PTC thermistor is reflected in the volt–ampere characteristic of Fig. 16.13. According to Ohm's law, a regular resistor with a near-zero TCR has a linear characteristic. A NTC thermistor has a positive curvature of the volt–ampere dependence. An implication of the negative TCR is that if such a thermistor is connected to a hard voltage source,[5] a self-heating due to Joule heat dissipation will result in resistance reduction. In turn, that will lead to a further increase in current and more heating. If the heat outflow from the NTC thermistor is restricted, a self-heating may eventually cause overheating and a catastrophic destruction of the device.

Because of positive TCRs, metals do not overheat when connected to hard voltage sources and behave as self-limiting devices. For instance, a filament in an incandescent lamp does not burn out because the increase in its temperature results in an increase in resistance, which limits current. This self-limiting (self-regulating) effect is substantially enhanced in the PTC thermistors. The shape of the volt–ampere characteristic indicates that in a relatively narrow temperature range, the PTC thermistor possesses a negative resistance; that is,

$$R_x = -\frac{V_x}{i}. \tag{16.39}$$

This results in the creation of an internal negative feedback which makes this device a self-regulating thermostat. In the region of negative resistance, any increase in voltage

[5] A hard voltage source means any voltage source having a near-zero output resistance and capable of delivering unlimited current without a change in voltage.

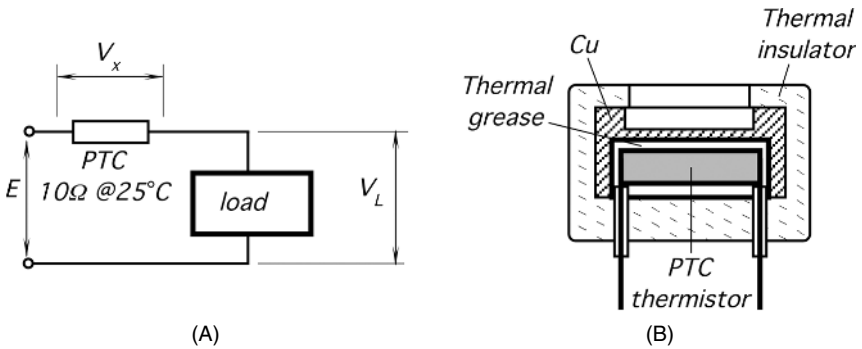


Fig. 16.14. Applications of PTC thermistors: (A) current-limiting circuit; (B) microthermostat.

across the thermistor results in heat production, which, in turn, increases the resistance and reduces heat production. As a result, the self-heating effect in a PTC thermistor produces enough heat to balance the heat loss on such a level that it maintains the device's temperature on a constant level T_0 (Fig. 16.12). That temperature corresponds to point x where the tangent to the curve has the highest value.

It should be noted that PTC thermistors are much more efficient when T_0 is relatively high (over 100°C) and their efficiency (the slope of the R–T curve near point x) drops significantly at lower temperatures. By their very nature, PTC thermistors are useful in the temperature range which is substantially higher than the operating ambient temperature.

There are several applications where the self-regulating effect of a PTC thermistor may be quite useful. We briefly mention four of them.

1. Circuit protection. A PTC thermistor may operate as a nondestructible (resettable) fuse in electric circuits, sensing excessive currents. Figure 16.14A shows a PTC thermistor connected in series with a power supply voltage E feeding the load with current i. The resistance of the thermistor at room temperature is quite low (typically from 10 to 140 Ω). The current i develops a voltage V_L across the load and a voltage V_x across the thermistor. It is assumed that $V_L \gg V_x$. Power dissipated by the thermistor, $P = V_x i$, is lost to the surroundings and the thermistor's temperature is raised above ambient by a relatively small value. Whenever either ambient temperature becomes too hot or load current increases dramatically (e.g., due to internal failure in the load), the heat dissipated by the thermistor elevates its temperature to a T_τ region where its resistance starts increasing. This limits further current increase. Under the shorted-load conditions, $V_x = E$ and the current i drops to its minimal level. This will be maintained until normal resistance of the load is restored and, it is said, the fuse resets itself. It is important to assure that $E < 0.9E_{\max}$, otherwise a catastrophic destruction of the thermistor may occur.
2. A miniature self-heating thermostat (Fig. 16.14B) for microelectronic, biomedical, chemical, and other suitable applications can be designed with a single PTC

thermistor. Its transition temperature must be appropriately selected. A thermostat consists of a dish, which is thermally insulated from the environment and thermally coupled to the thermistor. Thermal grease is recommended to eliminate a dry contact. The terminals of the thermistor are connected to a voltage source whose value may be estimated from

$$E \geq 2\sqrt{\delta(T_\tau - T_a)R_{25}}, \tag{16.40}$$

where δ is the heat dissipation constant which depends on thermal coupling to the environment and T_a is ambient temperature. The thermostat's set point is determined by the physical properties of the ceramic material (Curie temperature), and due to internal thermal feedback, the device reliably operates within a relatively large range of power-supply voltages and ambient temperatures. Naturally, the ambient temperature must be always less than T_τ.

3. Time delay circuits can be created with the PTC thermistors because of a relatively long transition time between the application of electric power in its heating and a low resistance point.
4. Flowmeter and liquid-level detectors which operate on the principle of heat dissipation can be made very simple with the PTC thermistors.

16.2 Thermoelectric Contact Sensors

Thermoelectric contact sensors are called *thermocouples* because at least two dissimilar conductors and two junctions (couples) of these conductors are needed to make a practical sensor. A thermocouple is a passive sensor. It generates voltage in response to temperature and does not require any external excitation power. The thermoelectric sensors belong to the class of the *relative* voltage-generating sensors, because the voltage produced depends on a *temperature difference* between two thermocouple junctions, in large part regardless of the absolute temperature of each junction. To measure temperature with a thermocouple, one junction will serve as a reference and its absolute temperature must be measured by a separate absolute sensor, such a thermistor, RTD, and so forth, or placed into a material that is in a state of a known reference temperature (Table 16.1). Section 3.9 of Chapter 3 provides a physical background for a better understanding of the thermoelectric effect and Table A.10 lists some popular thermocouples which are designated by letters originally assigned by the Instrument Society of America (ISA) and adopted by an American standard in ANSI MC 96.1. A detailed description of various thermocouples and their applications can be found in many excellent texts—for instance, Refs. [1], [9], and [10]. The most important recommendations for the use of these sensors are summarized as follows:

Type T: Cu (+) versus constantan (−) are resistant to corrosion in moist atmosphere and are suitable for subzero temperature measurements. Their use in air in an oxidizing environment is restricted to 370°C (700°F) due to the oxidation of the copper thermoelement. They may be used to higher temperatures in other atmospheres.

Type J: Fe (+) versus constantan (−) are suitable in vacuum and in oxidizing, reducing, or inert atmospheres over the temperature range of 0–760°C (32–1400°F). The rate of oxidation in the iron thermoelement is rapid above 540°C (1000°F), and the use of heavy-gauge wires is recommended when long life is required at the higher temperatures. This thermocouple is not recommended for use below the ice point because rusting and embrittlement of the iron thermoelement make its use less desirable than Type T.

Type E: 10% Ni/Cr (+) versus constantan (−) are recommended for use over the temperature range −200°C to 900°C (−330°F to 1600°F) in oxidizing or inert atmospheres. In reducing atmospheres, alternately oxidizing or reducing atmospheres, marginally oxidizing atmospheres, and in vacuum, they are subject to the same limitations as Type K. These thermocouples are suitable to subzero measurements because they are not subject to corrosion in atmospheres with a high moisture content. They develop the highest electromotive force (e.m.f.) per degree of all the commonly used types and are often used primarily because of this feature (see Fig. 3.36 of Chapter 3).

Type K: 10% Ni/Cr (+) versus 5%Ni/Al/Si (−) are recommended for use in an oxidizing or completely inert atmosphere over a temperature range of −200°C to 1260°C (−330°F to 2300°F). Due to their resistance to oxidation, they are often used at temperatures above 540°C. However, Type K should not be used in reducing atmospheres, in sulfurous atmospheres, and in a vacuum.

Types R and S: Pt/Rh (+) versus Pt (−) are recommended for continuous use in oxidizing or inert atmospheres over the temperature range 0–1480°C (32–2700°F).

Type B: 30% Pt/Rh (+) versus 6%Pt/Rh (−) are recommended for continuous use in oxidizing or inert atmospheres over the range 870–1700°C (1000–3100°F). They are also suitable for short-term use in a vacuum. They should not be used in reducing atmospheres in those containing metallic or nonmetallic vapors. They should never be directly inserted into a metallic primary protecting tube or well.

16.2.1 Thermoelectric Law

For practical purposes, an application engineer must be concerned with three basic laws which establish the fundamental rules for proper connection of the thermocouples. It should be stressed, however, that an electronic interface circuit must always be connected to two *identical* conductors. These conductors may be formed from one of the thermocouple loop arms. That arm is broken to connect the metering device to the circuit. The broken arm is indicated as material A in Fig. 16.15A.

> Law No. 1: A thermoelectric current cannot be established in a homogeneous circuit by heat alone.

This law provides that a nonhomogeneous material is required for the generation of the Seebeck potential. If a conductor is homogeneous, regardless of the temperature distribution along its length, the resulting voltage is zero. The junction of two dissimilar conductors provide a condition for voltage generation.

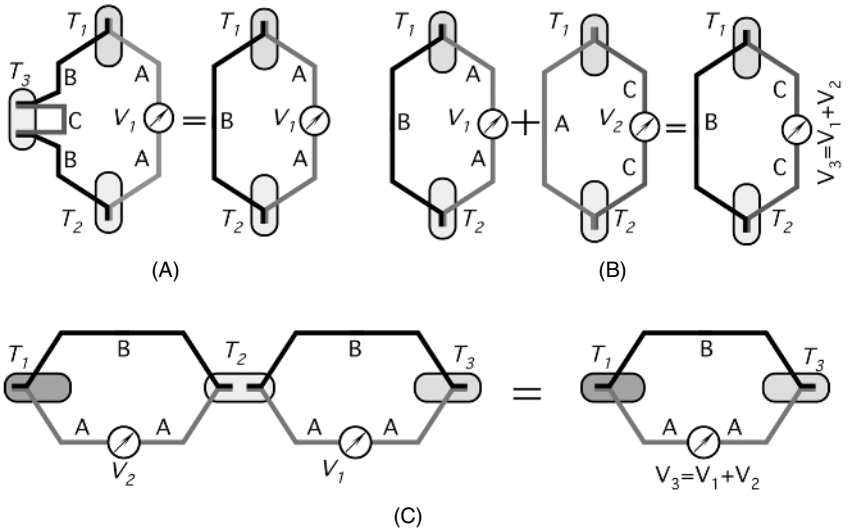


Fig. 16.15. Illustrations for the Laws of Thermocouples.

Law No. 2: The algebraic sum of the thermoelectric forces in a circuit composed of any number and combination of dissimilar materials is zero if all junctions are at a uniform temperature.

The law provides that an additional material C can be inserted into any arm of the thermoelectric loop without affecting the resulting voltage V_1 as long as both additional joints are at the same temperature (T_3 in Fig. 16.15A). There is no limitation on the number of inserted conductors, as long as both contacts for each insertion are at the same temperature. This implies that an interface circuit must be attached in such a manner as to assure a uniform temperature for both contacts. Another consequence of the law is that thermoelectric joints may be formed by any technique, even if an additional intermediate material is involved (such as solder). The joints may be formed by welding, soldering, twisting, fusion, and so on without affecting the accuracy of the Seebeck voltage. The law also provides a rule of *additive materials* (Fig. 16.15B): If thermoelectric voltages (V_1 and V_2) of two conductors (B and C) with respect to a reference conductor (A) are known, the voltage of a combination of these two conductors is the algebraic sum of their voltages against the reference conductor.

Law No. 3: If two junctions at temperatures T_1 and T_2 produce Seebeck voltage V_2, and temperatures T_2 and T_3 produce voltage V_1, then temperatures T_1 and T_3 will produce $V_3 = V_1 + V_2$ (Fig. 16.15C).

This is sometimes called the law of intermediate temperatures. The law allows us to calibrate a thermocouple at one temperature interval and then to use it at another interval. It also provides that extension wires of the same combination may be inserted into the loop without affecting the accuracy.

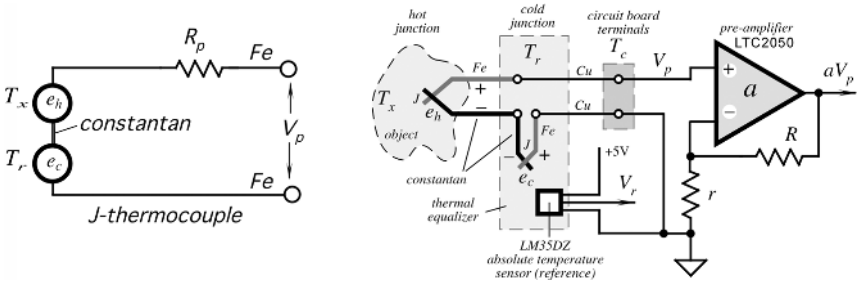


Fig. 16.16. Use of a thermocouple: (A) equivalent circuit of a thermocouple; (B) front end of a thermometer with a semiconductor reference sensor (LM35DZ).

Laws 1–3 provide for numerous practical circuits where thermocouples can be used in a great variety of combinations. They can be arranged to measure the average temperature of an object, to measure the differential temperature between two objects, and to use other than thermocouple sensors for the reference junctions and so forth.

It should be noted that thermoelectric voltage is quite small and the sensors, especially with long connecting wires, are susceptible to various transmitted interferences. A general guideline for the noise reduction can be found in Section 5.9 of Chapter 5. To increase the output signal, several thermocouples may be connected in series, while all reference junctions and all measuring junctions are maintained at the respective temperatures. Such an arrangement is called a *thermopile* (like piling up several thermocouples). Traditionally, the reference junctions are called *cold* and the measuring junctions are called *hot*.

Figure 16.16A shows an equivalent circuit for a thermocouple and a thermopile. It consists of a voltage source and a serial resistor. The voltage sources represent the *hot* (e_b) and *cold* (e_c) Seebeck potentials and the combined voltage V_p has a magnitude which is function of a temperature differential. The terminals of the circuit are assumed to be fabricated of the same material—iron in this example.

16.2.2 Thermocouple Circuits

In the past, thermocouples were often used with a cold junction immersed into a reference melting ice bath to maintain its temperature at 0°C (thus, the "cold" junction name). This presents serious limitations for many practical uses. The second and third thermoelectric laws allow for a simplified solution. A "cold" junction can be maintained at any temperature, including ambient, as long as that temperature is precisely known. Therefore, a "cold" junction is thermally coupled to an additional temperature sensor which does not require a reference compensation. Usually, such a sensor is either thermoresistive or a semiconductor.

Figure 16.16B shows the correct connection of a thermocouple to an electronic circuit. Both the "cold" junction and the reference sensor must be positioned in an intimate thermal coupling. Usually, they are imbedded in a chunk of copper. To

avoid dry contact, thermally conductive grease or epoxy should be applied for better thermal tracking. A reference temperature detector in this example is a semiconductor sensor LM35DZ manufactured by National Semiconductor, Inc. The circuit has two outputs: one for the signal representing the Seebeck voltage V_p and the other for the reference signal V_r. The schematic illustrates that connections to the circuit board input terminals and then to the amplifier's noninverting input and to the ground bus are made by the same type of wire (Cu). Both board terminals should be at the same temperature T_c; however, they do not necessarily have to be at the "cold" junction temperature. This is especially important for the remote measurements, where the circuit board temperature may be different from the reference "cold" junction temperature T_r.

For computing the temperature from a thermocouple sensor, two signals are essentially required. The first is a thermocouple voltage V_p and the other is the reference sensor output voltage V_r. These two signals come from different types of sensor and therefore are characterized by different transfer functions. A thermopile in most cases may be considered linear with normalized sensitivity α_p (V/K), whereas the reference sensor sensitivity is expressed according to its nature. For example, a thermistor's sensitivity α_r at the operating temperature T is governed by Eq. (16.21) and has dimension Ω/K. There are several practical ways of processing the output signals. The most precise method is to measure these signals separately, then compute the reference temperature T_r according to the reference sensor's equation, and compute the gradient temperature Δ from a thermocouple voltage V_p as

$$\Delta = T_x - T_r = \frac{V_p}{\alpha_p}. \tag{16.41}$$

Finally, add the two temperatures Δ and T_p to arrive at the measured absolute temperature T_x. A value of sensitivity (α_p) can be found from Table A.10.

For a relatively narrow reference temperature range, instead of adding up the temperatures, voltages from the reference sensor and the thermocouple can be combined instead. Because α_r and α_p are very much different, a scaling circuit must be employed. Figure 16.17 illustrates a concept of adding up voltages from a thermocouple and a thermistor (reference sensor) to obtain a combined output signal V_c. When adding up the voltages, the thermocouple amplifier gain should be selected to

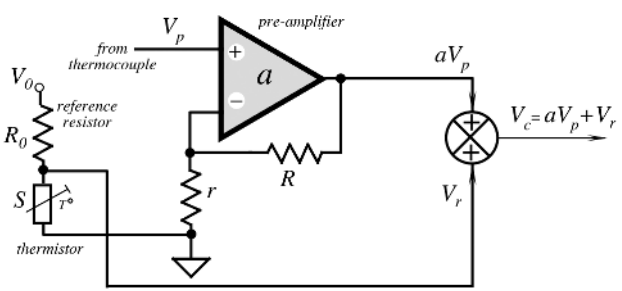


Fig. 16.17. Combining thermopile and thermistor signals.

match the temperature sensitivities of voltages V_p and V_r, or in other words, to satisfy condition

$$a\alpha_p = \alpha_r. \tag{16.42}$$

It is preferable to select $R_0 = S_0$ [S_0 is the thermistor resistance at the calibrating temperature T_0 (in Kelvin), for example at $T_0 = 298.15$ K (25°C) or in the middle of the operating temperature range]. With Eq. (16.21) in mind and after differentiating voltage V_r, we arrive at the amplifier's gain:

$$a = \frac{V_0}{\alpha_p}\frac{\beta}{T_0^2}\frac{R_0 S_0}{(R_0 + S_0)^2} \approx \frac{V_0}{4\alpha_p}\frac{\beta}{T_0^2}, \tag{16.43}$$

where V_0 is a constant voltage and β is the thermistor's characteristic temperature. The measured temperature can be computed from one of the corresponding equations found in Table 16.3, depending on the thermistor model used. When a particular thermistor model is selected, temperature is computed from a virtual thermistor's resistance S_c that first is derived from the combined voltage V_c as

$$S_c = R_0 \frac{V_c}{V_0 - V_c}. \tag{16.44}$$

16.2.3 Thermocouple Assemblies

A complete thermocouple sensing assembly generally consists of one or more of the following: a sensing element assembly (the junction), a protective tube (ceramic or metal jackets), a thermowell (for some critical applications, these are drilled solid bar stocks which are made to precise tolerances and are highly polished to inhibit corrosion), and terminations (contacts which may be in the form of a screw type, open type, plug and jack disconnect, military-standard-type connectors, etc.). Some typical thermocouple assemblies are shown in Fig. 16.18. The wires may be left bare or given electrical isolators. For the high-temperature applications, the isolators may be of a fish-spine or ball ceramic type, which provide sufficient flexibility. If thermocouple wires are not electrically isolated, a measurement error may occur. Insulation is affected adversely by moisture, abrasion, flexing, temperature extremes, chemical attack, and nuclear radiation. A good knowledge of particular limitations of insulating materials is essential for accurate and reliable measurement. Some insulations have a natural moisture resistance. Teflon, polyvinyl chloride (PVC), and some forms of polyimides are examples of this group. With the fiber-type insulations, moisture protection results from impregnating with substances such as wax, resins, or silicone compounds. It should be noted that only one-time exposure to ultraextreme temperatures cause evaporation of the impregnating materials and loss of protection.

The moisture penetration is not confined to the sensing end of the assembly. For example, if a thermocouple passes through hot or cold zones, condensation may produce errors in the measurement, unless adequate moisture protection is provided.

The basic types of flexible insulation for elevated temperature usage are fiber glass, fibrous silica, and asbestos (which should be used with proper precaution due

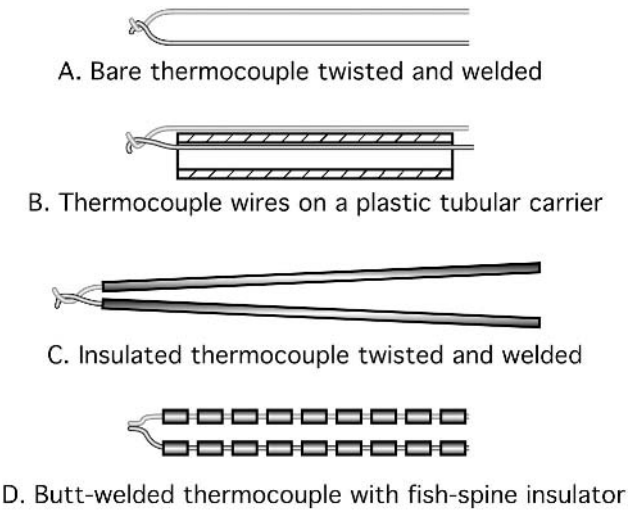

A. Bare thermocouple twisted and welded
B. Thermocouple wires on a plastic tubular carrier
C. Insulated thermocouple twisted and welded
D. Butt-welded thermocouple with fish-spine insulator

Fig. 16.18. Some thermocouple assemblies.

to health hazard). In addition, thermocouples must be protected from atmospheres that are not compatible with the alloys. Protecting tubes serve the double purpose of guarding the thermocouple against mechanical damage and interposing a shield between the wires and the environment. The protecting tubes can be made of carbon steels (up to 540°C in oxidizing atmospheres), stainless steel (up to 870°C), ferric stainless steel (AISI 400 series), and high-nickel alloys (Nichrome,[6] Inconel,[7] etc.) (up to 1150°C in oxidizing atmospheres).

Practically all base–metal thermocouple wires are annealed or given a "stabilizing heat treatment" by the manufacturer. Such treatment generally is considered sufficient, and seldom is it found advisable to further anneal the wire before testing or using. Although a new platinum and platinum–rhodium thermocouple wire as sold by some manufacturers is annealed already, it has become a regular practice in many laboratories to anneal all Type R, S, and B thermocouples, whether new or previously used, before attempting an accurate calibration. This is accomplished usually by heating the thermocouple electrically in air. The entire thermocouple is supported between two binding posts, which should be close together, so that the tension in the wires and stretching while hot are kept at a minimum. The temperature of the wire is conveniently determined with an optical pyrometer. Most of the mechanical strains are relieved during the first few minutes of heating at 1400–1500°C.

Thin-film thermocouples are formed by bonding junctions of foil metals. They are available in a free-filament style with a removable carrier and in a matrix style with a sensor embedded in a thin laminated material. The foil having a thickness in the order of 5 μm (0.0002 in.) gives an extremely low mass and thermal capacity. Thin

[6] Trademark of the Driver-Harris Company.

[7] Trademark of the International Nickel Company.

flat junctions may provide intimate thermal coupling with the measured surface. Foil thermocouples are very fast (a typical thermal time constant is 10 ms) and can be used with any standard interface electronic apparatuses. While measuring temperature with sensors having small mass, thermal conduction through the connecting wires always must be taken into account. Because of a very large length-to-thickness ratio of the film thermocouples (on the order of 1000), heat loss via wires is usually negligibly small.

To attach a film thermocouple to an object, several methods are generally used. Among them are various cements and flame or plasma-sprayed ceramic coatings. For ease of handling, the sensors often are supplied on a temporary carrier of polyimide film which is tough, flexible, and dimensionally stable. It is exceptionally heat resistant and inert. During the installation, the carrier can be easily peeled off or released by application of heat. The free foil sensors can be easily brushed into a thin layer, to produce an ungrounded junction. While selecting cements, care must be taken to avoid corrosive compounds. For instance, cements containing phosphoric acid are not recommended for use with thermocouples having copper in one arm.

16.3 Semiconductor P-N Junction Sensors

A semiconductor p-n junction in a diode and a bipolar transistor exhibits quite a strong thermal dependence [11]. If the forward-biased junction is connected to a constant-current generator (Fig. 16.19A) (Section 5.3.1 of Chapter 5), the resulting voltage becomes a measure of the junction temperature (Fig. 16.20). A very attractive feature of such a sensor is its high degree of linearity. This allows a simple method of calibration using just two points to define a slope (sensitivity) and an intercept.

The current-to-voltage equation of a p-n junction diode can be expressed as

$$I = I_0 \exp\left(\frac{qV}{2kT}\right), \tag{16.45}$$

where I_0 is the saturation current, which is a strong function of temperature. It can be shown that the temperature-dependent voltage across the junction can be expressed as

$$V = \frac{E_g}{q} - \frac{2kT}{q}(\ln K - \ln I), \tag{16.46}$$

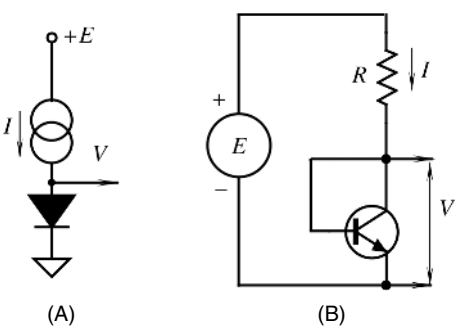


Fig. 16.19. Voltage-to-temperature dependence of a forward-biased semiconductor junction under constant-current conditions.

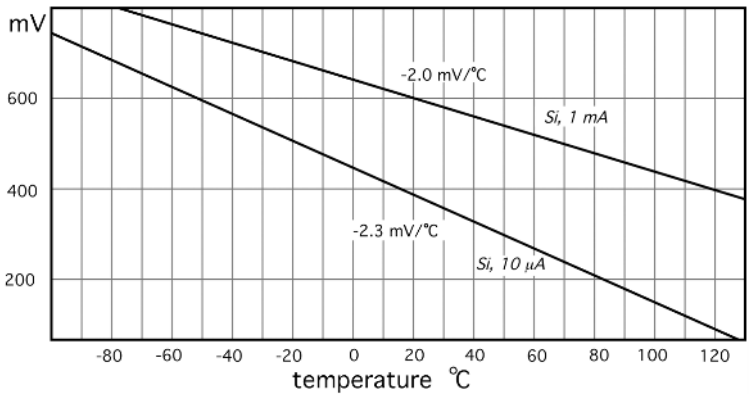


Fig. 16.20. Forward-biased p-n junction temperature sensors: (A) diode; (B) diode-connected transistor.

where E_g is the energy band gap for silicon at 0 K (absolute zero), q is the charge of an electron, and K is a temperature-independent constant. It follows from Eq. (16.46) that when the junction is operated under constant-current conditions, the voltage is linearly related to the temperature and the slope is given by

$$b = \frac{dV}{dT} - \frac{2k}{q}(\ln K - \ln I). \tag{16.47}$$

Typically, for a silicon junction operating at 10 μA, the slope (sensitivity) is approximately −2.3 mV/°C and it drops to about −2.0 mV/°C for a 1-mA current. Any diode or junction transistor can be used as a temperature sensor. A practical circuit for the transistor used as a temperature sensor is shown in Fig. 16.19B. A voltage source E and a stable resistor R is used instead of a current source. Current though the transistor is determined as

$$I = \frac{E - V}{R}. \tag{16.48}$$

It is recommended to use current on the order of $I = 100$ μA; therefore for $E = 5$ V and $V \approx 0.6$ V, the resistance $R = (E - V)/I = 44$ kΩ. When the temperature increases, the voltage V drops, which results in a minute increase in current I. According to Eq. (16.47), this causes some reduction in sensitivity which, in turn, is manifested as nonlinearity. However, the nonlinearity may be either small enough for a particular application or it can be taken care of during the signal processing. This makes a transistor (diode) temperature sensor a very attractive device for many applications, due to its simplicity and very low cost. Figure 16.21 shows an error curve for the temperature sensors made with the PN100 transistor operating at 100 μA. It is seen that the error is quite small, and for many practical purposes, no linearity correction is required.

A diode sensor can be formed in a silicon substrate in many monolithic sensors which require temperature compensation. For instance, it can be diffused into a mi-

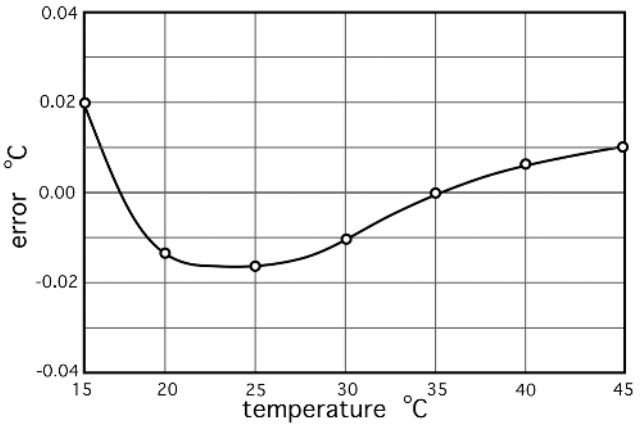


Fig. 16.21. An error curve for a silicon transistor (PN100) as a temperature sensor.

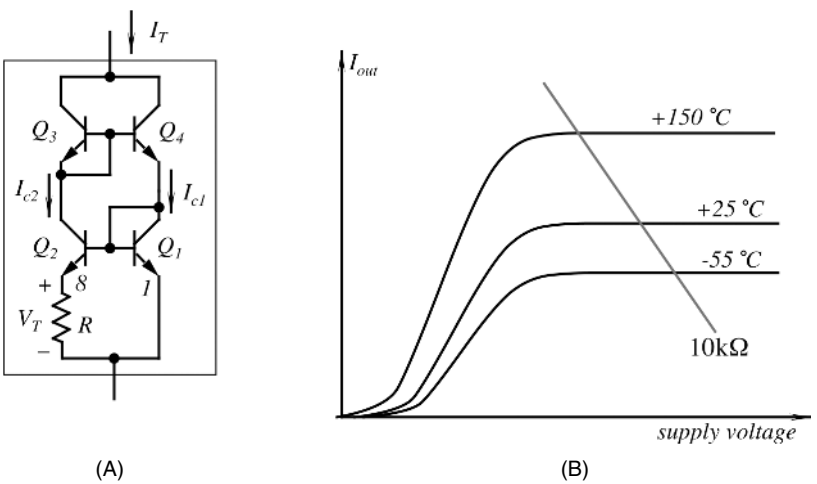


Fig. 16.22. Simplified circuit for a semiconductor temperature sensor (A) and current-to-voltage curves (B).

cromachined membrane of a silicon pressure sensor to compensate for temperature dependence of piezoresistive elements.

An inexpensive yet precision semiconductor temperature sensor may be fabricated by using fundamental properties of transistors to produce voltage which is proportional to absolute temperature (in Kelvin). That voltage can be used directly or it can be converted into current [12]. The relationship between base–emitter voltage (V_{be}) and collector current of a bipolar transistor is the key property to produce a linear semiconductor temperature sensor. Figure 16.22A shows a simplified circuit where Q_3 and Q_A form the so-called current mirror. It forces two equal currents $I_{C1} = I$ and $I_{C2} = I$ into transistors Q_1 and Q_2. The collector currents are determined by resistor

R. In a monolithic circuit, transistor Q_2 is actually made of several identical transistors connected in parallel, (e.g., eight). Therefore, the current density in Q_1 is eight times higher than that of each of transistors Q_2. The difference between base–emitter voltages of Q_1 and Q_2 is

$$\Delta V_{\text{be}} = V_{\text{be1}} - V_{\text{be2}} = \frac{kT}{q} \ln\left(\frac{rI}{I_{\text{ceo}}}\right) - \frac{kT}{q} \ln\left(\frac{I}{I_{\text{ceo}}}\right) = \frac{kT}{q} \ln r, \tag{16.49}$$

where r is a current ratio (equal to 8 in our example), k is the Boltzmann constant, q is the charge of an electron, and T is the temperature (in K). Currents I_{ceo} are the same for both transistors. As a result, a current across resistor R produces voltage $V_T = 179$ μV/K, which is independent of the collector currents. Therefore, the total current through the sensor is

$$I_T = 2\frac{V_T}{R} = \left(2\frac{k}{qR} \ln r\right) T, \tag{16.50}$$

which for current ratio $r = 8$ and resistance $R = 358$ Ω produces a linear transfer function $I_T / T = 1$ μA/K.

Figure 16.22B shows current-to-voltage curves for different temperatures. Note that the value in parentheses in Eq. (16.50) is constant for a particular sensor design and may be precisely trimmed during the manufacturing process for a desired slope I_T / T. The current I_T may be easily converted into voltage. If, for example, a 10-kΩ resistor is connected in series with the sensor, the voltage across that resistor will be a linear function of absolute temperature.

The simplified circuit of Fig. 16.22A will work according to Eqs. (16.49) and (16.50) only with perfect transistors ($\beta = \infty$). Practical monolithic sensors contain many additional components to overcome limitations of the real transistors. Several companies produce temperature sensors based on this principle. Examples are LM35 from National Semiconductors (voltage output circuit) and AD590 from Analog Devices (current output circuit).

Figure 16.23 shows a transfer function of a LM35Z temperature sensor which has a linear output internally trimmed for the Celsius scale with a sensitivity of 10 mV/°C. The function is quite linear where the nonlinearity error is confined within ±0.1°C. The function can be modeled by

$$V_{\text{out}} = V_0 + aT, \tag{16.51}$$

where T is the temperature in degrees Celsius. Ideally, V_0 should be equal to zero; however part-to-part variations of its value may be as large as ±10 mV, which corresponds to an error of 1°C. The slope a may vary between 9.9 and 10.1 mV/°C.

16.4 Optical Temperature Sensors

Temperature can be measured by contact and noncontact methods. The noncontact instruments are generally associated with the infrared optical sensors that we covered in Sections 3.12.3 of Chapter 3, 4.9 of Chapter 4, and 14.6 of Chapter 14. A need for the noncontact temperature sensors exists when the measurement must be done

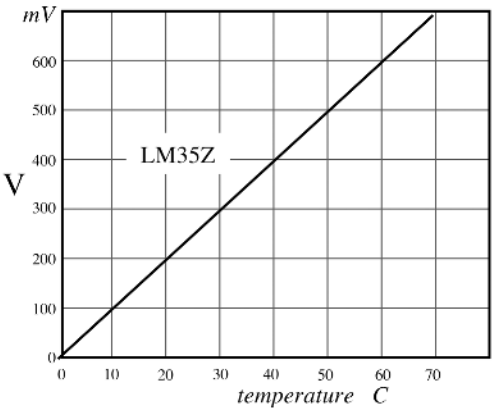


Fig. 16.23. A typical transfer function of a LM35DZ semiconductor temperature sensor; (Courtesy of National Semiconductors, Inc.).

quickly. Also, they are needed for determining temperatures at tough hostile environments when very strong electrical, magnetic, or electromagnetic fields or very high voltages make measurements either too susceptible to interferences or too dangerous for the operator. Also, there are situations when it is just difficult to reach an object during a routine measurement. In addition to the infrared methods of temperature measurements, there are sensors that are contact by nature but still use photons as carriers of thermal information.

16.4.1 Fluoroptic Sensors

These sensors rely on the ability of a special phosphor compound to give away a fluorescent signal in response to light excitation. The compound can be directly painted over the measured surface and illuminated by an ultraviolet (UV) pulse while observing the afterglow. The shape of the response afterglow pulse is function of temperature. The decay of the response pulse is highly reproducible over a wide temperature range [13,14]. As a sensing material, magnesium fluoromagnetite activated with tetravalent manganese is used. This is phosphor, long known in the lighting industry as a color corrector for mercury vapor street lamps, prepared as a powder by a solid-state reaction at approximately 1200°C. It is thermally stable, relatively inert, and benign from a biological standpoint, and insensitive to damage by most chemicals or by prolonged exposure to ultraviolet UV radiation. It can be excited to fluoresce by either UV or blue radiation. Its fluorescent emission is in the deep red region, and the fluorescent decay is essentially exponential.

To minimize cross-talk between the excitation and emission signals, they are passed through the bandpass filters, which reliably separate the related spectra (Fig. 16.24A). The pulsed excitation source, a xenon flash lamp, can be shared among a number of optical channels in a multisensor system. The temperature measurement is made by measuring the rate of decay of the fluorescence, as shown in Fig. 16.24B; that is, a temperature is represented by a time constant τ which drops fivefold over the

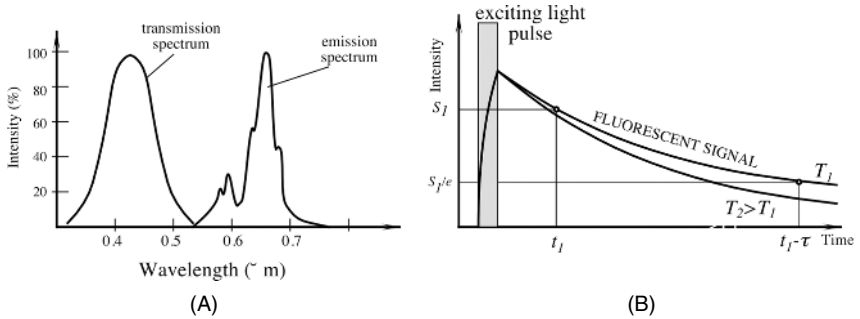


Fig. 16.24. Fluoroptic method of temperature measurement: (A) spectral responses of the excitation and emission signals; (B) exponential decay of the emission signal for two temperatures (T_1 and T_2); e is the base of natural logarithms and τ is a decay time constant. (Adapted from Ref. [13].)

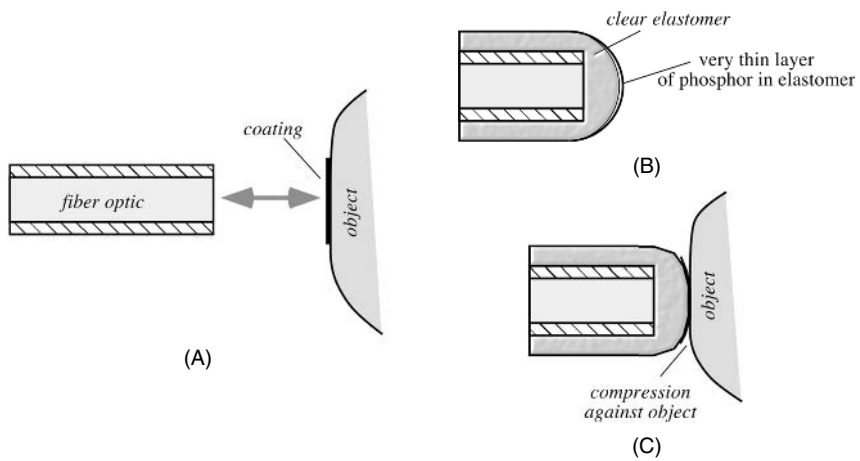


Fig. 16.25. Placement of a phosphor compound in the fluoroptic method: (A) on the surface of an object; (B and C) on the tip of the probe. (Adapted from Ref. [13].)

temperature range from −200°C to +400°C. The measurement of time is usually the simplest and most precise operation that can be performed by an electronic circuit; thus, temperature can be measured with a good resolution and accuracy—about ±2°C over the range without calibration.

Because the time constant is independent of excitation intensity, a variety of designs is possible. For instance, the phosphor compound can be directly coated onto the surface of interest and the optic system can take measurement without a physical contact (Fig. 16.25A). This makes possible the continuous temperature monitoring without disturbing a measured site. In another design, a phosphor is coated on the tip of a pliable probe which can form a good contact area when brought in contact with the object (Figs. 16.25B and 16.25C).

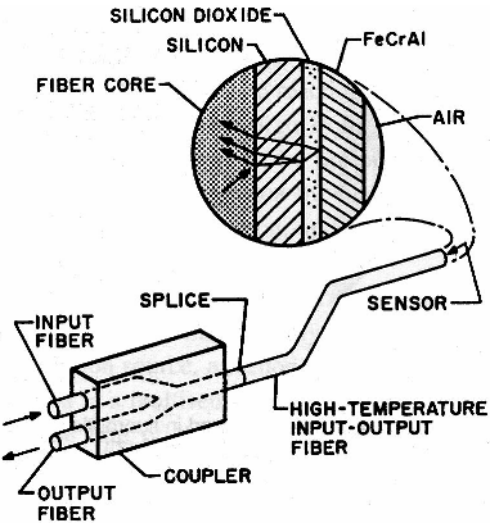


Fig. 16.26. A schematic of a thin-film optical temperature sensor.

16.4.2 Interferometric Sensors

Another method of optical temperature measurement is based on the modulation of light intensity by interfering two light beams. One beam is a reference, and the other travels through a temperature-sensitive medium and is somewhat delayed depending on temperature. This results in a phase shift and a subsequent extinction of the interference signal. For temperature measurement, a thin layer of silicon [15,16] can be used because its refractive index changes with temperature, thus modulating a light travel distance.

Figure 16.26 shows a schematic of a thin-film optical sensor. The sensor was fabricated by sputtering three layers onto the ends of the step-index multimode fibers with 100-μm core diameters and 140-μm cladding diameters [17]. The first layer is silicone, then silicon dioxide. The FeCrAl layer on the end of the probe prevents oxidation of the underlying silicon. The fibers can be used up to 350°C, however, much more expensive fibers with gold-buffered coatings can be used up to 650°C. The sensor is used with a LED source operating in the range of 860 nm and a micro-optic spectrometer.

16.4.3 Thermochromic Solution Sensor

For biomedical applications, where electromagnetic interferences may present a problem, a temperature sensor can be fabricated with the use of a thermochromic solution [18], such as cobalt chloride ($CoCl_2 \cdot 6H_2O$). The operation of this sensor is based on the effect of a temperature dependence of a spectral absorption in the visible range of 400–800 nm by the thermochromic solution (Fig. 16.27A). This implies that

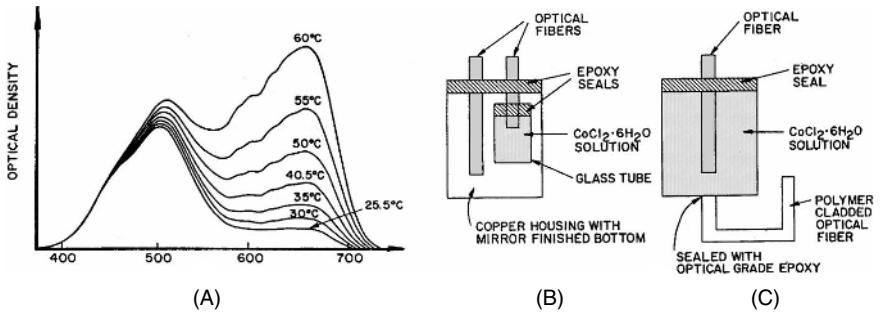


Fig. 16.27. A thermochromic solution sensor: (A) absorption spectra of the cobalt chloride solution; (B) reflective fiber coupling; (C) transmissive coupling. (From Ref. [18].)

the sensor should consist of a light source, a detector, and a cobalt chloride solution, which is thermally coupled with the object. Two possible designs are shown in Figs. 16.27B and 16.27C, where transmitting and receiving optical fibers are coupled through a cobalt chloride solution.

16.5 Acoustic Temperature Sensor

Under extreme conditions, temperature measurement may become a difficult task. These conditions include a cryogenic temperature range, high radiation levels inside nuclear reactors, and so forth. Another unusual condition is the temperature measurement inside a sealed enclosure with a known medium, in which no contact sensors can be inserted and the enclosure in not transmissive for the infrared radiation. Under such unusual conditions, acoustic temperature sensors may come in quite handy. An operating principle of such a sensor is based on a relationship between temperature of the medium and speed of sound. For instance, in dry air at a normal atmospheric pressure, the relationship is

$$v \approx 331.5\sqrt{\frac{T}{273.15}}\frac{\text{m}}{\text{s}}, \tag{16.52}$$

where v is the speed of sound and T is the absolute temperature.

An acoustic temperature sensor (Fig. 16.28) is composed of three components: an ultrasonic transmitter, an ultrasonic receiver, and a gas-filled hermetically sealed tube. The transmitter and receiver are ceramic piezoelectric plates which are acoustically decoupled from the tube to assure sound propagation primarily through the enclosed gas, which, in most practical cases, is dry air. Alternatively, the transmitting and receiving crystals may be incorporated into a sealed enclosure with a known content whose temperature has to be measured; that is, an intermediate tube in not necessarily required in cases where the internal medium, its volume, and mass are held constant. When a tube is used, care should be taken to prevent its mechanical deformation and

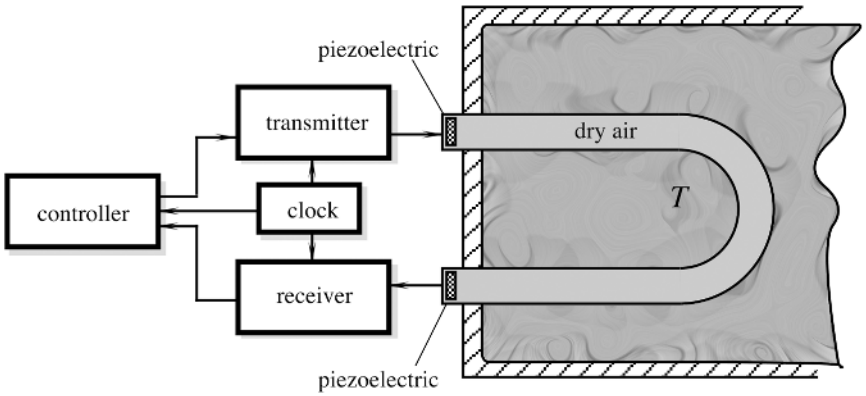


Fig. 16.28. An acoustic thermometer with an ultrasonic detection system.

loss of hermeticity under the extreme temperature conditions. A suitable material for the tube is Invar.

The clock of low frequency (near 100 Hz) triggers the transmitter and disables the receiver. The piezoelectric crystal flexes, transmitting an ultrasonic wave along the tube. The receiving crystal is enabled before the wave arrives at its surface and converts it into an electrical transient, which is amplified and sent to the control circuit. The control circuit calculates the speed of sound by determining the propagation time along the tube. Then, the corresponding temperature is determined from the calibration numbers stored in a look-up table. In another design, the thermometer may contain only one ultrasonic crystal which alternatively acts either as a transmitter or as a receiver. In that case, the tube has a sealed empty end. The ultrasonic waves are reflected from the end surface and propagate back to the crystal, which, before the moment of the wave arrival, is turned into a reception mode. An electronic circuit [19] converts the received pulses into a signal which corresponds to the tube temperature.

A miniature temperature sensor can be fabricated with the surface-acoustic-wave (SAW) and plate-wave (PW) techniques (see Chapter 11). The idea behind such a sensor is in the temperature modulation of some mechanical parameters of a time-keeping element in the electronic oscillator [20,21]. This leads to the change in the oscillating frequency. In effect, such an integral acoustic sensor becomes a direct converter of temperature into frequency. A typical sensitivity is in the range of several kilohertz per degree Kelvin.

16.6 Piezoelectric Temperature Sensors

The piezoelectric effect, in general, is a temperature-dependent phenomenon. Thus, a temperature sensor based on the variability of the oscillating frequency of a quartz crystal can be designed. Because the quartz is an anisotropic medium, the resonant frequency of a plate is highly dependent on the crystallographic orientation of the

plate—the so-called angle of cut. Thus, by selecting a cut, a negligibly small temperature sensitivity may be achieved (AT- and BT-cuts), or just the opposite—a cut with pronounced temperature dependence may be selected. The temperature dependence of the resonant frequency may be approximated by a third-order polynomial:

$$\frac{\Delta f}{f_0} = a_0 + a_1 \Delta T + a_2 \Delta T^2 + a_3 \Delta T^3 \tag{16.53}$$

where ΔT and Δf are the temperature and frequency shifts respectively, f_0 is the calibrating frequency, and a are the coefficients. The first utilization of temperature dependence was made in 1962 by utilizing a nonrotated Y-cut crystal [22]. A very successful development of a linear temperature coefficient cut (LC) was made by Hewlett- Packard [23]. The second- and third-order coefficients had been eliminated by selecting a doubly-rotated Y-cut. The sensitivity (a_1) of the sensor is 35 ppm/°C and the operating temperature range is from −80°C to 230°C with a calibration accuracy of 0.02°C. With the advent of microprocessors, linearity became a less important factor, and more sensitive, yet somewhat nonlinear quartz temperature sensors had been developed by using a slightly singly rotated Y-cut ($Q = -4$°C) with sensitivity of 90 ppm/°C [24] and by utilizing a tuning-fork resonators in flexural and torsional modes [25,26].

It should be noted that thermal coupling of the object of measurement with the oscillating plate is always difficult and, thus, all piezoelectric temperature sensors have a relatively slow response as compared with thermistors and thermoelectrics.

References

1. Benedict, R. P. *Fundamentals of Temperature, Pressure, and Flow Measurements*, 3rd ed. John Wiley & Sons, New York, 1984.
2. Callendar, H. L. On the practical measurement of temperature. *Phil. Trans. R. Soc. London* 178, 160, 1887.
3. Sapoff, M. Thermistor thermometers. In: *The Measurement, Instrumentation and Sensors Handbook*. J.G. Webster, ed., CRC Press, Boca Raton, FL, 1999, pp. 32.25–32.41.
4. Fraden, J. A two-point calibration of negative temperature coefficient thermistors. *Rev. Sci. Instrum.* 71(4), 1901–1905, 2000.
5. Steinhart, J.S. and Hart, S.R. *Deep Sea Res.*, 15, 497, 1968.
6. Mangum, B.W. *Rev. Sci. Instrum.* 54(12), 1687, 1983.
7. Sapoff, M., Siwek, W.R., Johnson, H.C., Slepian, J., and Weber, S. In: *Temperature. Its Measurement and Control in Science and Industry*. J.E. Schooley, ed. American Institute of Physics, New York, 1982, Vol. 5, p. 875.
8. *Keystone NTC and PTC Thermistors. Catalogue* Keystone Carbon Company, St. Marys, PA, 1984.
9. Caldwell, F.R. *Thermocouple Materials*. NBS monograph 40. National Bureau of Standards, Washington, DC, 1962.

10. *Manual on the Use of Thermocouples in Temperature Measurement*, 4th ed. ASTM Manual Series: MNL: 12-93. ASTM, Philadelphia, 1993.
11. Sachse, H. B. *Semiconducting Temperature Sensors and Their Applications*. Wiley–Interscience, New York, 1975.
12. Timko, M. P. A two terminal IC temperature transducer. *IEEE J. Solid- State Circuits*. SC-11, 784–788, 1976.
13. Wickersheim, K.A. and Sun, M.H. Fluoroptic thermometry. *Med. Electron.*, 84–91, Febr. 1987.
14. Fernicola, V.C. et al. Investigations on exponential lifetime measurements for fluorescence thermometry. *Rev. Sci. Instrum.* 71(7), 2938–2943, 2000.
15. Schultheis, L., Amstutz, H., and Kaufmann, M. Fiber-optic temperature sensing with ultrathin silicon etalons. *Opt. Lett.* 13(9), 782–784, 1988.
16. Wolthuis, R., A., Mitchell, G.L., Saaski, E., Hartl, J.C., and Afromowitz, M.A. Development of medical pressure and temperature sensors employing optical spectral modulation. *IEEE Trans. Biomed. Eng.* 38(10), 974–981, 1991.
17. Beheim, G., Fritsch, K., and Azar, M.T. A sputtered thin film fiber optic temperature sensor. *Sensors Magazine*, 37–43, Jan. 1990.
18. Hao, T. and Lui, C.C. An optical fiber temperature sensor using a thermochromic solution. *Sensors Actuators A* 24, 213–216, 1990.
19. Williams, J. Some techniques for direct digitization of transducer outputs, In: *Linear Technology Application Handbook*. Linear Technology Inc., 1990.
20. Venema, A., et al. Acoustic-wave physical-electronic systems for sensors. *Fortschritte der Akustik der 16. Deutsche Arbeitsgemeinschaft für Akustik*, pp. 1155–1158, 1990.
21. Vellekoop, M.J., et al. All-silicon plate wave oscillator system for sensor applications. Proc. IEEE Ultrasonic Symposium, 1990.
22. Smith, W.L. and Spencer, L.J. Quartz crystal thermometer for measuring temperature deviation in the 10-3 to 10-6 °C range. *Rev. Sci. Instrum.* 268–270, 1963.
23. Hammond, D.L. and Benjaminson, A. Linear quartz thermometer. *Instrum. Control Syst.* 38, 115, 1962.
24. Ziegler H. A low-cost digital sensor system. *Sensors Actuators*, 5, 169–178, 1984.
25. Ueda, T., Kohsaka, F., Iino, T., and Yamazaki, D. Temperature sensor utilizing quatz tuning fork resonator. Proc. 40th Ann. Freq. Control Symp., 1986, pp. 224–229.
26. EerNisse E.P., and Wiggins, R.B. A resonator temperature transducer with no activity dips. Proc. 40th Ann. Freq. Control Symp., 1986, pp. 216–223.

17

Chemical Sensors[1]

Chemical sensors respond to stimuli produced by various chemicals or chemical reactions. These sensors are intended for the *identification* and *quantification* of chemical species (including both liquid and gaseous phases; solid chemical sensors are not common).

In science and research, chemical sensors are used in many areas from atmospheric monitoring of pollutant emissions to detection of explosives. These sensors are used routinely to characterize gas samples from laboratory experiments and to track the migration of hazardous chemical spills in soils at field sites. New applications include tracking/locating insect pest infestations such as termites by their characteristic off-gassing from cellulose digestion and the monitoring of the menstrual cycles of cattle (to improve effectiveness of artificial insemination).

In industry, chemical sensors are used for process and quality control during plastics manufacturing and in the production of foundry metals where the amount of diffused gases affects metal characteristics such as brittleness. They are used for environmental monitoring of workers to control their exposure to dangers and limit health risks. Chemical sensors find many new applications as *electronic noses* and are being used to test and control food spoilage, the distribution of pesticides in agricultural applications, and to grade beverages.

In medicine, chemical sensors are used to determine patient health by monitoring oxygen and trace gas content in the lungs and in blood samples. These sensors are often used for breathalyzers to test for blood alcohol levels and as indicators of the digestion problems of patients.

In the military, chemical sensors are used to detect fuel dumps and airborne chemical warfare agents. Liquid chemical sensors are used to manage training base operations by carefully monitoring groundwater contamination. Combinations of liquid and gas sensors are used in experimental military applications to monitor toxics produced from refineries and nuclear plants to verify compliance with weapons treaties.

[1] This chapter is written in collaboration with Dr. Michael C. Vogt (Argonne National Laboratory).

17.1 Chemical Sensor Characteristics

Most chemical sensors can be described using criteria and characteristics general to all sensors such as stability, repeatability, linearity, hysteresis, saturation, response time, and span (see Chapter 2), but two characteristics are unique and meaningful as applied to chemical detection. Because chemical sensors are used both for identification and quantification, they need to be both selective and sensitive to a desired target species in a mixture of chemical species.

Selectivity describes the degree to which a sensor responds to only the desired target species, with little or no interference from nontarget species. *Sensitivity* describes the minimal concentrations and concentration changes (then referred to as *resolution*) that can be successfully and repeatedly sensed by a device. Note that for the sensors described in the previous chapters, the term "sensitivity" is often used as a synonym of "slope" when the transfer function of a sensor is linear. For the chemical sensors, sensitivity is the synonym of resolution. This is a characteristic that other sensors, like pressure and temperature, are rarely concerned with.

Therefore, one of the most important functions in the evaluation of a chemical sensor's performance is the qualification of its selectivity. It is common practice to evaluate the response of a sensor only for increasing the values of activity (concentration) to the primary target species. This is mainly due to the fact that it is more convenient to prepare a continuously broad range of test concentrations by adding increasing amounts of a concentrated (pure) primary species to the background sample than vice versa. An absolutely selective sensor really does not exist and there is commonly some interference present.

17.2 Specific Difficulties

The difficulty of developing chemical sensors versus other sensors (such as temperature, pressure, humidity, etc.) is that *chemical reactions change the sensor*, often in a way that is nonreversible. For example, electrochemical cells employing liquid electrolytes (material that conducts electrical current via charged ions, not electrons) lose a small amount of electrolytes with each measurement, requiring that the electrolyte be replenished eventually or have carbonic acid forming at the gate–membrane interface and etching components in chemical field-effect transistor (FET) sensors.

Also, unlike pressure or temperature sensors which have comparatively few conditions under which they need to be modeled to operate, chemical sensors are often exposed to nearly unlimited numbers of chemical combinations. This introduces interference responses, contamination from acid attack, or, for porous film devices, the sorption of species that cannot be removed (such as silicone on zirconia sensors), altering their surface area and effectively changing their calibrated behavior.

For the ceramic bead-type catalytic hydrocarbon sensors, bulk platinum electrodes and heating elements begin to evaporate at elevated (1000°C) temperatures, limiting their life spans and their usefulness for long-term continuous monitoring [1]. This evaporation rate is even higher in the presence of combustible gases. The loss of the

platinum metal results in a change in the resistance of the wire that introduces offset error into the sensor reading, and it leads to early burnout of the heating platinum coil.

Chemical poisoning can affect many sensors like the catalytic bead devices where silicone and ethyl lead bind to the sensing element, inhibiting the oxidation of the hydrocarbon species and producing an inaccurate, false low reading. Filters are commonly used with any chemical sensor if it is to be subjected to an environment containing a characteristic poison. Judicious selection of the filter material is required to eliminate only the poisoning agent without an associated reduction in the target analyte (the chemical species being exposed to the sensor).

Surface-acoustical-wave (SAW) devices that use species-selective adsorptive films can be poisoned *mechanically* by species that adsorb but which do not desorb returning the mass of the device back to its original (calibrated) state. Similarly, gas-selective coatings on fiber-optic devices also may be poisoned by nonremovable species, permanently reducing the optical reflectance and indicating a false positive.

Another problem unique to chemical sensors is the significant chemical reaction changes that occur throughout the concentration levels. Reactive hydrocarbon devices (metal-oxide devices, voltammetric devices, etc.) require mixtures near stoichiometric (balanced chemical reactions) so that required minimal levels of both target analyte hydrocarbons and needed oxygen are available to feed the measurement reaction. If the hydrocarbon levels are too high (or better stated as the accompanying oxygen levels are too low), then only a fraction of the hydrocarbons will react producing a false-negative reading again.

17.3 Classification of Chemical-Sensing Mechanisms

All chemical sensing can be classified by the actual indicator phenomena employed for sensing, and they also can be classified by the measurement strategy employed. We will separate chemical sensors into two major groups, *direct (simple)* and *indirect (complex)*, and will also distinguish between *chemically* reactive devices and *physically* reactive devices in each group (Fig. 17.1).

Direct chemical sensors utilize any of a variety of chemical reaction phenomena that directly affect a measurable electrical characteristic such as resistance, potential,

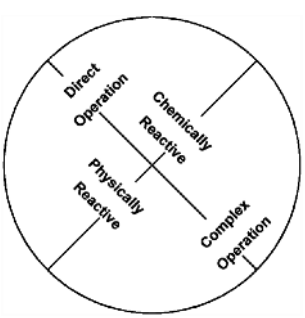


Fig. 17.1. Direct versus complex and chemically versus physically reactive groupings.

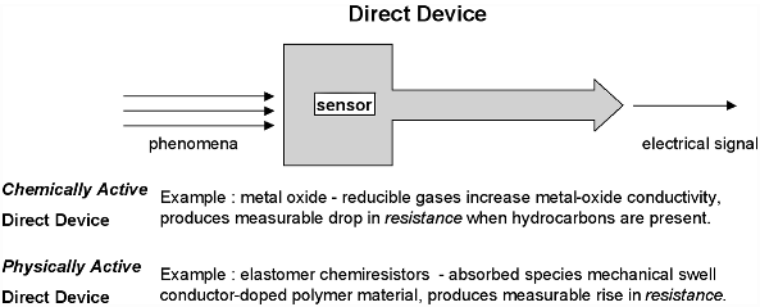

Direct Device
sensor
phenomena
electrical signal
Chemically Active
Direct Device
Example : metal oxide - reducible gases increase metal-oxide conductivity, produces measurable drop in resistance when hydrocarbons are present.
Physically Active
Direct Device
Example : elastomer chemiresistors - absorbed species mechanical swell conductor-doped polymer material, produces measurable rise in resistance.

Fig. 17.2. Direct devices.

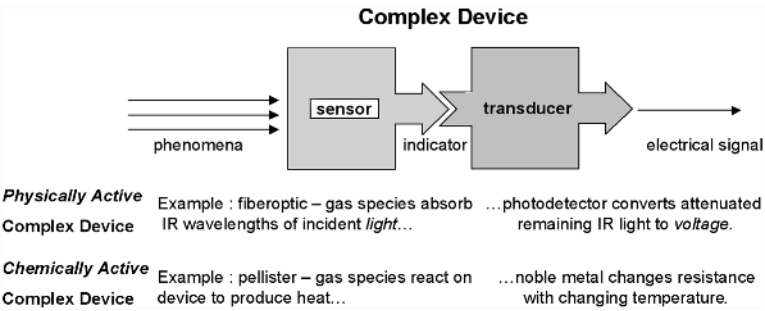


Fig. 17.3. Complex devices.

current, or capacitance (Fig. 17.2). These devices require some sort of electrical signal conditioning, but no *transducing* (converting the sensor phenomena from one form of energy to another). *Complex* devices (Fig. 17.3) employ chemistry-influenced phenomena that do not directly affect an electrical characteristic and will require some form of transducing to obtain electrical signal in order to interface with common measurement electronics. Nondirect phenomena include physical shape change, frequency shifts, modulation of light, temperature or produced heat change, and even mass change.

Some of the simplest chemical-sensor designs require that the sensing element *chemically* react with the analyte to effect a measurable change in the indicator (phenomena) or signal. This often adversely influences the device and introduces stability problems. The chemically reactive devices suffer when there is incomplete reversibility, when there is depletion or consumption of the sensor/analyte chemicals (electrochemical cells use up electrolyte and some electrodes get consumed), or when there is no species-specific reaction (including interference from other species).

Physical chemical sensors do not require a chemical reaction to take place, but isolate and employ a *physical* reaction to indicate the presence of a chemical species. These devices regularly demonstrate less drift and better stability than true chemically reactive devices, but often at the cost of significant additional instrumentation and slower reaction times.

17.4 Direct Sensors

Direct chemical sensors that affect the electrical characteristics of a sensing element can be separated into categories by the characteristic that they affect. *Conductometric* devices affect the resistance or impedance of the sensing element, *amperometric* devices affect the measurable electrical or electronic current passing through the sensing element, and *potentiometric* devices affect the electrical potential or voltage across some pair of electrodes. Through circuitry, these characteristics can be readily converted from one characteristic to another to simplify interfacing. There are a wide variety of chemical-sensing phenomena that employ direct sensing.

17.4.1 Metal-Oxide Chemical Sensors

Metal-oxide gas sensors (such as tin dioxide, SnO_2) have been popular since the late 1960s [2]. They are simple rugged devices that perform reasonably well with relatively simple electronics support. Bulk metal oxides have electrical properties that change in the presence of reducible gases such as methyl mercaption (CH_3SH) and ethyl alcohol (C_2H_5OH). When a metal-oxide crystal such as SnO_2 is heated at a certain high temperature in air, oxygen is adsorbed on the crystal surface and a surface potential is formed that inhibits electron flow. When the surface is exposed to reducible gases, the surface potential decreases and conductivity measurably increases.

The relationship between the film's electrical resistance and a given reducible gas' concentration is described by the following empirical equation:

$$R_s = A[C]^{-\alpha}, \tag{17.1}$$

where R_s is the sensor electrical resistance, A is a constant specific for a given film composition, C is the gas concentration, and α is the characteristic slope of the R_s curve for that material and expected gas.

Metal-oxide devices change the resistivity as a function of the presence of reducible gases, and as such, they require an additional electronic circuit to operate. A typical arrangement is to design the sensor as one leg in a common Wheatstone bridge circuit so that the changing resistance can be detected as an unbalancing of the potential drops observed across the bridge circuit (Fig. 17.4A). The negative temperature coefficient (NTC) thermistor (temperature sensor; see Chapter 16) with a linearizing parallel resistor is required to adjust the bridge balance point according to the sensor's temperature.

Because the sensor behaves as a resistance whose value is controlled by gas species and gas concentration, the voltage drop across it is proportional to its resistance and a plot of voltage drop versus gas concentration is recorded. The response signal from the sensors is linear when plotted on logarithmic charts (Fig. 17.4B). The slopes and offsets of the curves produced by different reducible gases allow them to be distinguished from each other and quantified within certain concentration ranges where the curves do not overlap [3]. Optionally, the rate of change of the conductivity may be used to differentiate gases and concentrations [4]. The bulk conductivity can drift for these devices, but the rate of change of that conductivity when driven by a pulsed input is more stable and reproducible.

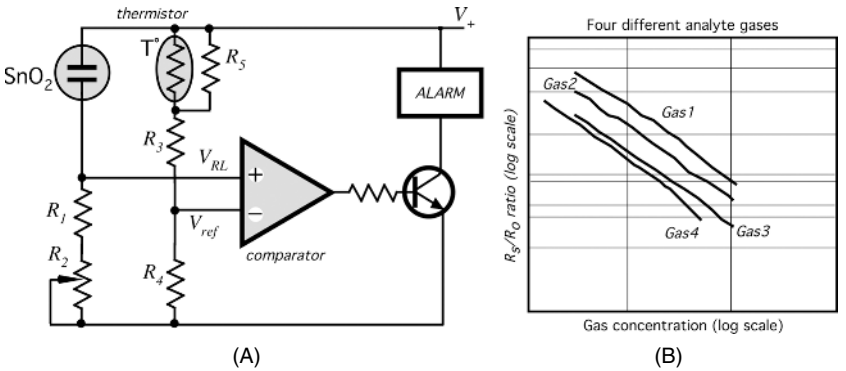


Fig. 17.4. SnO_2 Wheatstone bridge circuit (A) and its response for different gases (B).

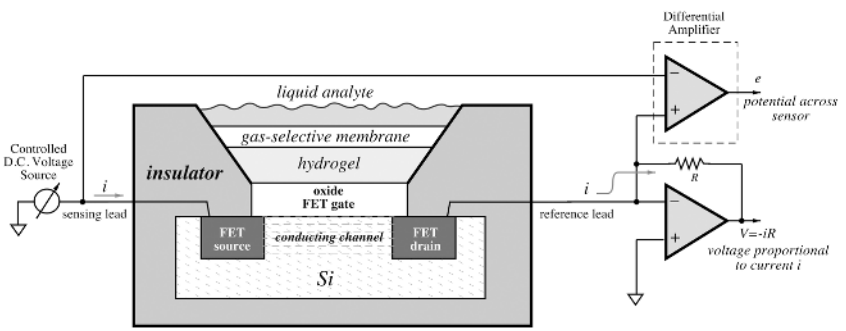


Fig. 17.5. Liquid ChemFET construction and electrical connection.

17.4.2 ChemFET

A chemFET is a chemical field-effect transistor that includes a gas-selective coating or series of coatings between its transistor gate and the analyte (Fig. 17.5). This chemical element gives the device a control input that modifies the source–drain conduction in relationship with selected chemical species. Different materials applied to the gate react with different chemical species (gases or liquids) and provide differentiation of species. ChemFETs can be used for detecting H_2 in air, O_2 in blood, some military nerve gases, NH_3, CO_2, and explosive gases [5].

As in a conventional FET, the chemFET is constructed using thin-film techniques and commonly employs a *p*-type silicon body with two n-type silicon diffusion regions (source and drain). This three-part system is covered with a silicon dioxide insulator layer separating a final top metal gate electrode above and between the source and drain. Operation involves applying a voltage, positive with respect to the silicon to the gate electrode. Electrons are attracted to the surface of the semiconductor forming a conducting channel between the source and the drain *n*-regions [6]. In fact, a chemFET is a chemically controlled conductor (resistor). Conductance of a chemFET is measured by a differential amplifier and is represented by the output

voltage e. To compute conductance, the current in the circuit is measured by the I/V converter with a reference resistor R.

Hydrogen-gas-sensing chemFETs use a palladium/nickel (Pd/Ni) film as their gates [7]. The improved, more stable, chemFETs used for liquid sensing employ a silver/sliver chloride hydrogel (Ag/AgCl) bridge between the silicon dioxide (SiO_2) gate and a selective membrane that separates the gate from the analyte (Fig. 17.5). The selective membrane is commonly polyvinyl chloride (PVC), polyurethane, silicone rubber, or polystyrene.

For an ion-selective chemFET the gate is replaced by or coated with a chemical-selective electrolyte or other semiconductor material. If the ion-sensitive material is ion penetrable, then the device is called a MEMFET, and if the membrane is ion impenetrable, it is called a SURFET. The chemical-selective gate material alters the potential at which the device begins to conduct and thus indicates the presence of specific chemical species. The devices are inherently small and low in power consumption. The gate coatings for the chemFET can be enzyme membranes (ENFET) or ion-selective membranes (ISFET). Ion-selective membranes produce a chemical sensor, and enzyme membranes can produce a biochemical sensor. The enzyme membrane is made from polyaniline (PANIE) and is, itself, created using a voltammetric electrochemical process to produce this organic semiconductor.

17.4.3 Electrochemical Sensors

The electrochemical sensors are the most versatile and better developed than any other chemical sensors. Depending on the operating mode, they are divided into sensors which measure voltage (*potentiometric*), those which measure electric current (*amperometric*), and those which rely on the measurement of conductivity or resistivity (*conductometric*). In all of these methods, special electrodes are used, where either a chemical reaction takes place or the charge transport is modulated by the reaction. A fundamental rule of an electrochemical sensor is that it always requires a closed circuit; that is, an electric current (either dc, or ac) must be able to flow in order to make a measurement. Because electric current flow essentially requires a closed loop, the sensor needs at least two electrodes, one of which is called a *return electrode*. It should be noted, however, that even if, in the potentiometric sensors, no flow of current is required for the voltage measurement, the loop still must be closed for the measurement of voltage.

The electrodes in these sensing systems are often made of catalytic metals such as platinum or palladium or they can be carbon-coated metals. Electrodes are designed to have a high surface area to react with as much of the analyte as possible, producing the largest measurable signal. Electrodes can be treated (modified) to improve their reaction rates and extend their working life spans. The *working* electrode (WE) is where the targeted chemical reactions take place (Fig. 17.6). The electrical signal is measured with respect to a *counter or auxiliary* electrode (AE) which is not intended to be catalytic, and in the case of three-electrode systems, a third *reference* electrode (RE) is employed to measure and correct for electrochemical potentials generated by each electrode and the electrolyte. The third electrode improves operation by correcting for error introduced by a polarization of the working electrode. Newer electrochem-

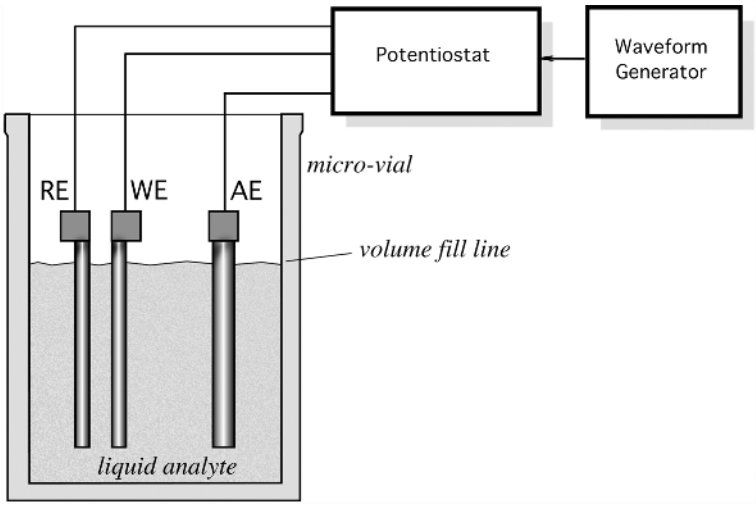


Fig. 17.6. Electrochemical-sensor electrode set.

ical sensors employ thick-film screen-printed electrode sets to make manufacturing simpler and more robust.

The electrolyte is a medium that carries charges using ions instead of electrons. This directly limits the reactions that can take place and is the first stage of lending selectivity to the electrochemical sensor. The sensor formed by this collection of electrodes and electrolytes is called an *electrochemical cell* and may be operated in several ways depending on the electrical characteristic (resistance, potential, current, capacitance, etc.) being observed. The more comprehensive measurements are captured using various forms of voltammetry discussed later in this chapter.

A simple liquid electrochemical sensor (cell) uses two electrodes immersed in an electrolyte solution. Gas analytes such as CO react at the working electrode and produces CO_2 and free electrons. Charges and charged species migrate to the other (counter) electrode where water is produced if oxygen is present. The reaction converts CO to CO_2. If the electrodes are connected in series to a resistor and the potential drop across the resistor is measured, it will be proportional to the current flowing, making it a function of analyte gas present.

17.4.4 Potentiometric Sensors

These sensors use the effect of the concentration on the equilibrium of the redox reactions occurring at the electrode–electrolyte interface in a electrochemical cell. An electrical potential may develop at this interface due to the redox reaction which takes place at the electrode surface, where Ox denotes the oxidant and Red denotes the reduced product [8]:

$$\mathrm{Ox} + \mathrm{Ze} = \mathrm{Red}. \tag{17.2}$$

This reaction occurs at one of the electrodes (cathodic reaction in this case) and is called a half-cell reaction. Under thermodynamical quasiequilibrium conditions, the Nernst equation is applicable and can be expressed as

$$E = E_0 + \frac{RT}{nF} \ln\left(\frac{C_0^*}{C_R^*}\right), \tag{17.3}$$

where C_0^* and C_R^* are concentrations of Ox and Red, respectively, n is the number of electrons transferred, F is the Faraday constant, R is the gas constant, T is the absolute temperature, and E_0 is the electrode potential at a standard state. In a potentiometric sensor, two half-cell reactions will take place simultaneously at each electrode. However, only one of the reactions should involve the sensing species of interest; the other half-cell reaction is preferably reversible, noninterfering, and known.

The measurement of the cell potential of a potentiometric sensor should be made under zero-current or quasiequilibrium conditions; thus, a very high-input-impedance amplifier (which is called an *electrometer*) is generally required. There are two types of electrochemical interface from the viewpoint of the charge transfer: ideally polarized (purely capacitive) and nonpolarized. Some metals (e.g., Hg, Au, Pt) in contact with solutions containing only an inert electrolyte (e.g., H_2SO_4) approach the behavior of the ideally polarized interface. Nevertheless, even in those cases, a finite charge-transfer resistance exists at such an interface and excess charge leaks across with the time constant given by the product of the double-layer capacitance and the charge-transfer resistance ($\tau = R_{ct}C_{dl}$).

An ion-selective membrane is the key component of all potentiometric ion sensors. It estab- lishes the reference with which the sensor responds to the ion of interest in the presence of various other ionic components in the sample. An ion-selective membrane forms a nonpolarized interface with the solution. A well-behaved membrane (i.e., one which is stable, reproducible, immune to adsorption and stirring effects, and also selective) has both high absolute and relative exchange-current density.

17.4.5 Conductometric Sensors

An electrochemical conductivity sensor measures the change in conductivity of the electrolyte in an electrochemical cell. An electrochemical sensor may involve a capacitive impedance resulting from the polarization of the electrodes and faradic or charge-transfer process.

In a homogeneous electrolytic solution, the conductance of the electrolyte, $G(\Omega^{-1})$, is inversely proportional to L, which is the segment of the solution along the electrical field, and directly proportional to A, which is the cross-sectional area perpendicular to the electric field [9]:

$$G = \frac{\rho A}{L}, \tag{17.4}$$

where $\rho(\Omega^{-1}\text{cm}^{-1})$ is the specific conductivity of the electrolyte and is related quantitatively to the concentration and the magnitude of the charges of the ionic species.

The equivalent conductance of the solution at any concentration, C in mol/L or any convenient units, is given by

$$\Lambda = \Lambda_0 - \beta C^{0.5}, \tag{17.5}$$

where β is a characteristic of the electrolyte and Λ_0 is the equivalent conductance of the electrolyte at an infinite dilution.

Measurement techniques of electrolytic conductance by an electrochemical conductivity sensor has remained basically the same over the years. Usually, a Wheatstone bridge (similar to Fig. 17.4) is used with the electrochemical cell (the sensor) forming one of the resistance arms of the bridge. However, unlike the measurement of the conductivity of a solid, the conductivity measurement of an electrolyte is often complicated by the polarization of the electrodes at the operating voltage. A faradic or charge-transfer process occurs at the electrode surfaces. Therefore, a conductivity sensor should be operated at a voltage where no faradic process could occur. Another important consideration is the formation of a double layer adjacent to each of the electrodes when a potential is imposed on the cell. This is described by the so-called Warburg impedance. Hence, even in the absence of the faradic process, it is essential to take into consideration the effect of the double layers during measurement of the conductance. The effect of the faradic process can be minimized by maintaining the high cell constant L/A of the sensor so that the cell resistance lies in the region between 1 and 50 kΩ. This implies using a small electrode surface area and large interelectrode distance. This, however, reduces the sensitivity of the Wheatstone bridge. Often the solution is in the use of a multiple-electrode configuration. Both effects of the double layers and the faradic process can be minimized by using a high-frequency low-amplitude alternating current. Another good technique would be to balance both the capacitance and the resistance of the cell by connecting a variable capacitor in parallel to the resistance of the bridge area adjacent to the cell.

17.4.6 Amperometric Sensors

An example of an amperometric chemical sensor is a Clark oxygen sensor which was proposed in 1956 [10,11]. The operating principle of the electrode is based on the use of an electrolyte solution contained within the electrode assembly to transport oxygen from an oxygen-permeable membrane to the metal cathode. The cathode current arises from a two-step, oxygen-reduction process that may be represented as

$$\begin{aligned} &O_2 + 2H_2O + 2e^- \rightarrow H_2O_2 + 2OH^- \\ &H_2O_2 + 2e^- \rightarrow 2OH^-. \end{aligned} \tag{17.6}$$

Figure 17.7A shows the membrane which is stretched across the electrode tip, allowing oxygen to diffuse through a thin electrolyte layer to the cathode. Both anode and cathode are contained within the sensor assembly, and no electrical contact is made with the outside sample. A first-order diffusion model of the Clark electrode is illustrated in Fig. 17.7B [11]. The membrane–electrolyte–electrode system is considered

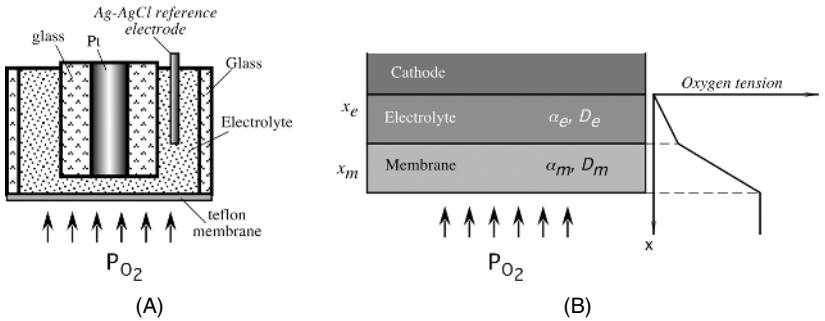


Fig. 17.7. Clark electrode (A) and the first-order one-dimensional model (B) of the oxygen tension distribution throughout the system. (Adapted from Ref. [11].)

to act as a one-dimensional diffusion system with the partial pressure at the membrane surface equal to the equilibrium partial pressure p_0 and that at the cathode equal to zero. It can be shown that the steady-state electrode current is given by

$$I \approx \frac{4Fa_m D_m p_0}{x_m}, \tag{17.7}$$

where A is the electrode area, αm is the solubility of oxygen in the membrane, F is the Faraday's constant, D_m is the diffusion constant, and x_m is the thickness of the membrane. It should be noted that the current is independent of the electrolyte thickness and diffusion properties. A Teflon® membrane is used as an oxygen-permeable film. We may define the sensor's sensitivity as a ratio of the current to the oxygen partial pressure:

$$S = \frac{I}{p_0}. \tag{17.8}$$

For example, if the membrane is 25 μm thick and the cathode area is 2×10^{-6} cm^2, then the sensitivity is approximately 10^{-12} A/mm Hg.

An enzymatic-type amperometric sensor can be built with a sensor capable of measuring the relative oxygen deficiency caused by the enzymatic reaction by using two Clark oxygen electrodes. The operating principle of the sensor is shown in Fig. 17.8. The sensor consists of two identical oxygen electrodes, where one (A) is coated with an active oxidize layer and the other (B) with an inactive enzyme layer. An example of the application is a glucose sensor, where inactivation can be carried out either chemically, by radiation, or thermally. The sensor is encapsulated into a plastic carrier with glass coaxial tubes supporting two Pt cathodes and one Ag anode. In the absence of the enzyme reaction, the flux of oxygen to these electrodes and, therefore, the diffusion-limiting currents are approximately equal to one another. When glucose is present in the solution and the enzymatic reaction takes place, the amount of oxygen reaching the surface of the active electrode is reduced by the amount consumed by the enzymatic reaction, which results in a current imbalance.

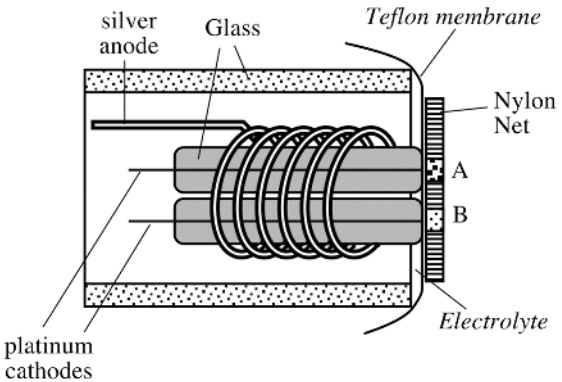


Fig. 17.8. Simplified schematic of an amperometric Clark oxygen sensor adapted for detecting glucose.

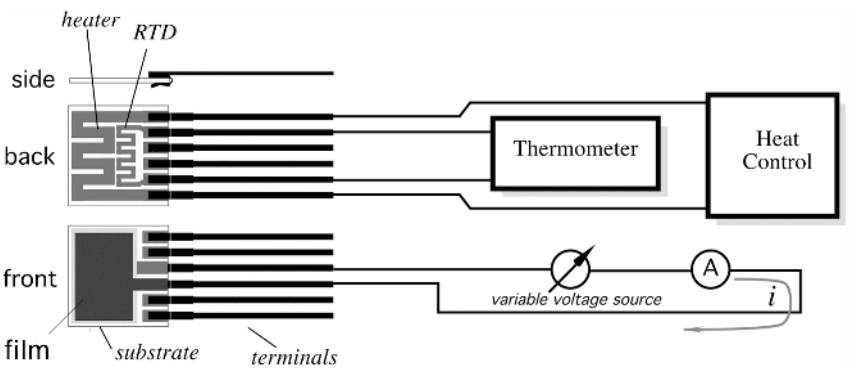


Fig. 17.9. Electrocatalytic gas microsensor fabricated on a ceramic substrate having a heating element and resistive temperature sensor (RTD) on one side and thick-film solid electrolyte on the other side.

17.4.7 Enhanced Catalytic Gas Sensors

Enhanced catalytic gas sensors are experimental devices that employ active measurement techniques coupled with fairly simple electrochemical cells [12]. The electrochemical cells are fabricated of ceramic–metallic films and provide the reaction environment for potentiometric and amperometric measurements. These sensors provide broad-spectrum responses allowing identification and quantification of a wide range of gases. The enhanced catalytic devices are divided into *electro*-enhanced catalytic devices (electrocatalytic) and *photo*-enhanced catalytic devices (photocatalytic).

The electrocatalytic devices employ a thick-film solid electrolyte electrochemical cell (Fig. 17.9).The cell is fabricated using screen-printing/firing techniques to produce a sandwich of ceramic–metallic (cermet) materials on a 625-µm-thick aluminum oxide substrate (Al_2O_3). The lower reference electrode measures approximately 15

μm thick and is made of platinum (Pt) bonded to nickel oxide (NiO). The upper sensing electrode measures approximately 5 μm thick and is made of platinum (Pt) sintered into a porous structure. The two electrodes are separated by a 25–30-μm-thick yttria-stabilized zirconia (YSZ) solid electrolyte. The final cross-sectional arrangement is Al_2O_3\Pt\Ni–NiO\YSZ\Pt. In the simple diffusion-driven mode, an electric potential is produced as a function of the natural log of the ratio of partial pressures of the gases on opposing faces of the sensor, as described by the Nernst equation [Eq. (17.3)]. This is a very limited sensing reaction, but is the one most used by common automotive oxygen sensors.

When the electrodes of the device are excited by an external driving potential, more complex and interesting chemical reactions are initiated. As a changing potential is applied, the gas species at the surface of the device will reduce or oxidize and release or capture free electrons [Eq. (17.2)]. This reaction affects the electrical current *i* passing through the film, which can be measured by an ammeter as a function of the gas species and applied potential. The current is affected by both the rate of change of the applied potential and the reactive Faradic current component [13]. By altering the time-based shape of the applied potential, these two signal components can be separated and better used to detect gas species. Because the reaction depends on temperature, a heater and temperature sensor are incorporated into the sensor to maintain temperature on a predetermined level.

The photocatalytic devices (Fig. 17.10) employ materials such as titanium dioxide (TiO_2) as a catalyst. TiO_2 dramatically enhances electro-oxidation reactions when it is exposed to wavelengths of ultraviolet (UV) light less than 320 nm. The photocatalytic sensors change resistance when exposed to the proper wavelength of UV light and a reactable gas species [14]. These devices can be used with a single excitation wavelength to simply detect the presence of a gas species by it gross resistance chance, or they can be coupled with several different UV light sources and doped TiO_2 films to change the reaction windows and improve speciation of gas analytes.

Both changing applied potentials and changing activation light sources can be used to excite the enhanced catalytic devices to a state where they are reactive to different chemical species. With the electrocatalytic devices, the gas species on the surface react at species-specific dissociation potentials. This attenuates or augments the electrical current passing through the device that is recorded as a spike or drop in measured current. Because of their requirement for advanced active measurement techniques, the enhanced catalytic sensors occupy a role between simple sensors and fully equipped instruments.

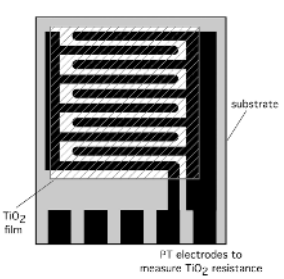


Fig. 17.10. Photocatalytic gas microsensor.

17.4.8 Elastomer Chemiresistors

Elastomer chemiresistors or polymer conductive composites (also *polymer conductors* or simply PCs) are polymer films that adsorb chemical species and swell, increasing resistance as a *physical* response to the presence of a chemical species. These can be used as chemical detectors but do not truly employ a *chemical* reaction. The polymers are designed and/or treated to attract subsets of chemicals providing a degree of speciation or selectivity. PCs have become commercially viable [15] as the sensing element inside of a more complex instrument. The PC sensors can respond to the presence of simple hydrocarbons like isopropyl alcohol in only a couple of seconds, whereas more complex oils may take 10–15 s. The PC element is not expected to be tolerant of corrosives, but barring exposure to such, it should have a life span of months in normal operation. The PC measurement strategy uses several differently treated PC elements to produce an array, and then it samples the array to produce a signature. The commercial instruments based on this technology can readily differentiate between compounds such as acetone and acetic acid, but they are not designed to be quantitative. These commercial instruments are complements to metal-oxide sensors in that they are rather insensitive to fixed gases like O_2, Cl_2, H_2, and NO that are commonly detected using metal-oxide devices. Unlike metal-oxide-based sensors, the PCs do not require the high controlled operating temperatures and consume significantly less power.

To detect the presence of a liquid, a sensor usually must be specific to that particular agent at a certain concentration; that is, it should be selective to the liquid's physical and/or chemical properties. An example of such a sensor is a resistive detector of hydrocarbon fuel leaks (originally devised in Bell Communication Research to protect buried telephone cables). A detector is made of silicone and carbon black composite. The polymer matrix serves as the sensing element and the conductive filler is used to achieve a relatively low volume resistivity, on the order of 10 Ω cm in the initial standby state. The composition is selectively sensitive to the presence of a solvent with a large solvent–polymer interaction coefficient [16]. Because the sensor is not susceptible to polar solvents such as water or alcohol, it is compatible with the underground environment. The sensor is fabricated in the form of a thin film with a very large surface/thickness ratio. Whenever the solvent is applied to the film sensor, the polymer matrix swells, resulting in the separation between conductive particles. This causes a conversion of the composite film from being a conductor to becoming an isolator with a resistivity on the order 10^9 Ω cm, or even higher. The response time for a film sensor is less than 1 s. The sensor returns to its normally conductive state when it is no longer in contact with the hydrocarbon fuel, making the device reusable.

17.5 Complex Sensors

Complex sensors involve chemical phenomena that change the state of an indicator as a function of some chemical reaction. The indicator can be a temperature change, an opacity change, an oscillation frequency change, and so forth. These indicators require another transducer to convert the changing indicator to an electrical output.

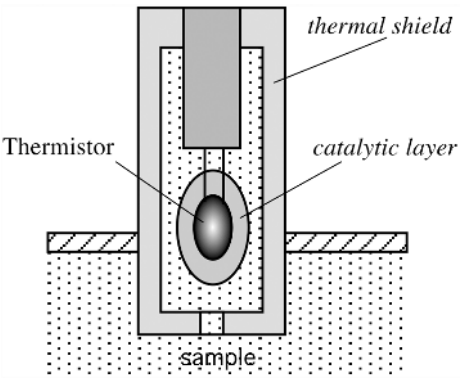


Fig. 17.11. Schematic diagram of a chemical thermal sensor.

17.5.1 Thermal Sensors

When the internal energy of a system changes, it is accompanied by an absorption or evolution of heat. This is called the first law of thermodynamics. Therefore, a chemical reaction which is associated with heat can be detected by an appropriate thermal sensor, such as those described in Chapter 16. These sensors operate on the basic principles which form the foundation of *microcalorimetry*. An operating principle of a thermal sensor is simple: A temperature probe is coated with a chemically selective layer. Upon the introduction of a sample, the probe measures the release of heat during the reaction between the sample and the coating.

A simplified drawing of such a sensor is shown in Fig. 17.11. It contains a thermal shield to reduce heat loss to the environment and a catalytic-layer-coated thermistor. The layer may be an enzyme immobilized into a matrix. An example of such a sensor is the enzyme thermistor using an immobilized oxidize (GOD). The enzymes are immobilized on the tip of the thermistor, which is then enclosed in a glass jacket in order to reduce heat loss to the surrounding solution. Another similar sensor with similarly immobilized bovine serum albumin is used as a reference. Both thermistors are connected as the arms of a Wheatstone bridge [17]. The temperature increase as a result of a chemical reaction is proportional to the incremental change in the enthalpy, dH:

$$dT = \frac{1}{C_p}\, dH, \tag{17.9}$$

where C_p is the heat capacity.

The chemical reaction in the coating is

$$\beta - \text{D-glucose} + H_2O + O_2 \overset{\text{GOD}}{\rightarrow} H_2O + \text{D-gluconic acid},\ \Delta H_1, \tag{17.10}$$

and

$$H_2O_2 \overset{\text{catalase}}{\rightarrow} \frac{1}{2}O_2 + H_2O,\ \Delta H_2, \tag{17.11}$$

where ΔH_1 and ΔH_2 are partial enthalpies, the sum of which for the above reaction is approximately -80 kJ/mol. The sensor responds linearly with the dynamic range, depending on the concentration of hydrogen peroxide (H_2O_2).

17.5.2 Pellister Catalytic Sensors

These sensors operate on the principle similar to thermal enzymatic sensors. Heat is liberated as a result of a catalytic reaction taking place at the surface of the sensor and the related temperature change inside the device is measured. On the other hand, the chemistry is similar to that of high-temperature conductometric oxide sensors. Catalytic gas sensors have been designed specifically to detect a low concentration of flammable gases in ambient air inside mines. These sensors often are called *pellistors* [8]. The platinum coil is imbedded in a pellet of ThO_2/Al_2O_3 coated with a porous catalytic metal: palladium or platinum. The coil acts as both the heater and the resistive temperature detector (RTD). Naturally, any other type of heating element and temperature sensor can be successfully employed. When the combustible gas reacts at the catalytic surface, the heat evolved from the reaction increases the temperature of the pellet and of the platinum coil, thus increasing its resistance. There are two possible operating modes of the sensor. One is isothermal, where an electronic circuit controls the current through the coil to maintain its temperature constant. In the nonisothermal mode, the sensor is connected as a part of a Wheatstone bridge whose output voltage is a measure of the gas concentration.

17.5.3 Optical Chemical Sensors

Optical sensors are based on the interaction of electromagnetic radiation with matter, which results in altering (modulating) some properties of the radiation. Examples of such modulations are variations in intensity, polarization, and velocity of light in the medium. The presence of different chemicals in the analyte affects which wavelengths of light are modulated. Optical modulation is studied by spectroscopy, which provides information on various microscopic structures from atoms to the dynamics in polymers. In a general arrangement, the monochromatic radiation passes through a sample (which may be gas, liquid, or solid), and its properties are examined at the output. Alternatively, the sample may respond with a secondary radiation (induced luminescence), which is also measured.

Chemiluminescence devices (reaction produces measurable light) phosphoresce when light hits them and that emission of light is an indication of chemical species presence. Nondispersive infrared (NDIR) absorbance involves the absorption of specific wavelengths of light and, when tuned through experimental methods, can be used for single-analyte target gases such as CO_2. *Spectroscopic* absorption optical sensors are useful for UV and IR wavelengths and can be used to target O_3 detection by producing a more complex absorbance signature versus a simple attenuation. In all strategies, the wavelength of the light source is routinely matched to the reactive energy of the optrode indicator to achieve a best possible electronic signal. The detection of the original and resultant light is obtained with a photodiode or photomultiplier tube.

Optical chemical sensors can be and are designed and built in a great variety of ways, which are limited only by the designer's imagination. Here, we will describe only one device just to illustrate how an optical sensor works. Figure 17.12 shows a simplified configuration of a CO_2 sensor [18]. It consists of two chambers which are illuminated by a common LED. Each chamber has metallized surfaces for better internal reflectivity. The left chamber has slots covered with a gas-permeable membrane.

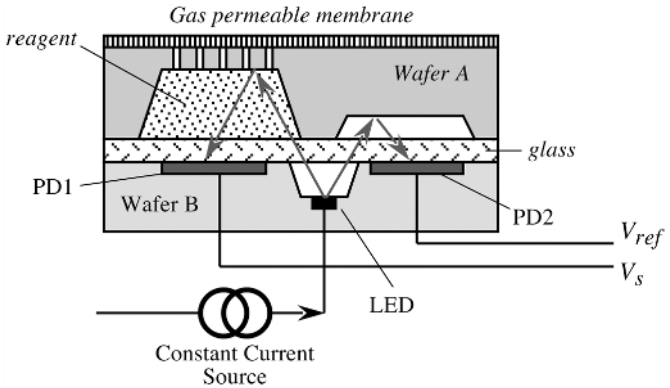


Fig. 17.12. Simplified configuration of an optical CO_2 sensor.

The slots allow CO_2 to diffuse into the chamber. The bottom parts of the chambers are made of glass. Both wafers, A and B, form optical waveguides. The test chamber is filled with a reagent, and the reference chamber is not. The sample part of the sensor monitors the optical absorbency of a pH indicator in a dilute solution, where the optical absorbency changes in accordance with the Beer–Lambert law:

$$I = I_0 \exp\left[-\alpha\,(\lambda, \mathrm{pH})\,\mathrm{dC}\right], \tag{17.12}$$

where I is the transmitted intensity, I_0 is the source intensity, a is the molar absorptivity, λ is the wavelength, C is the concentration, and d is the optical path length.

Ambient CO_2 equilibrates with the bicarbonate ion buffer system in the reagent, as it is done in the traditional Severinghaus–Stow CO_2 electrode. Equilibrium among CO_2, H_2CO_3, and HCO_3 produces a change in the pH of the solution. The solution contains a pH indicator Chlorophenol Red, which exhibits a sharp, nearly linear change in the optical absorbency at 560 nm from pH 5 to pH 7. The buffer concentration can be selected to exhibit pH changes in the range for partial CO_2 pressures from 0 to 140 torr. Because the buffer pH varies linearly with the log of the partial pressure of carbon dioxide (pCO_2), changes in optical absorbency can also be expected to vary linearly with the log of pCO_2.

The LED common for both halves of the sensor transmits light through the pH-sensitive sample to a test photodiode (PD1). The second photodiode (PD2) is for reference purposes to negate variations in the light intensity of the LED. For temperature stability, the sensor should operate in a thermally stable environment.

Fiber-optic chemical sensors (Fig. 17.13) use a chemical reagent phase to alter the amount or wavelength of light reflected by, absorbed by, or transmitted through a fiber waveguide (see also Fig. 4.17A of Chapter 4). A fiber-optic sensor typically contains three parts: a source of incident (pilot) light, an optrode, and a transducer (detector), to convert the changing photonic signal to an electrical signal. It is the optrode that contains the reagent phase membrane or indicator whose optical properties are affected by the analyte [19].

The location of the reagent, and the specific optical characteristic that is affected by it, vary from one type of optical sensor to another. Simple polymer-coated fibers

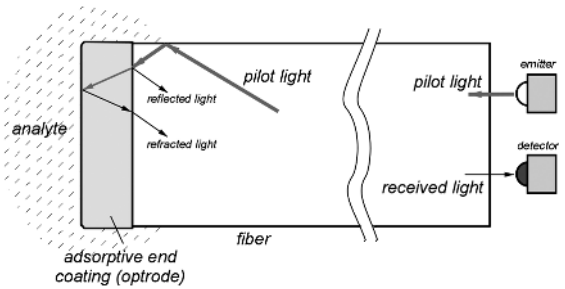


Fig. 17.13. Fiber-optic gas sensor.

coat the polished lens end of a glass fiber with a reagent that absorbs incident light. Coating the cladding of a fiber instead of its polished end affects the reflection and refraction of the light. This is referred to as evanescent wave sensing. Whereas the glass optical fiber is rugged and in many cases chemically resistant, the coating or indicator is not and becomes the weak component in the system [20].

Differential designs (to isolate all but reaction of interest) are often employed to split the original incoming light source and pass one through the reagent area while the other is unaltered. The two optical paths are either multiplexed to a single detector (transducer) or fed to different transducers to produce a difference signal used for sensing.

17.5.4 Mass Detector

Chemical sensors that utilize the very small mass change from adsorbed chemical molecules to alter mechanical properties of a system are referred to as *mass*, *gravimetric*, or *microbalance sensors*. These are physically active devices, as no chemical reaction takes place. Measurement of microscopic amount of mass cannot be accomplished by using conventional balances; the quantity of the material is just too small. So the oscillating sensor had been invented. Sometimes, it is called an acoustic gravimetric sensor because it operates at ultrasonic frequencies. The idea behind the oscillating sensor is the shift in the resonant frequency of a piezoelectric crystal when an additional mass is deposited on its surface. A piezoelectric quartz oscillator resonates with a frequency which, depending on the circuit, is called either a series (fr) or a parallel (far) resonant (see Fig. 7.39B of Chapter 7). Either frequency is a function of the crystal mass and shape. In a simplified manner, the acoustic gravimetric sensor may be described as an oscillating plate whose natural frequency depends on its mass. Adding material to that mass would shift the frequency and thus can be measured by electronic means:

$$\frac{\Delta f}{f_o} = S_m \Delta m \tag{17.13}$$

where f_0 is the unloaded natural oscillating frequency, Δf is the frequency shift: ($\Delta f = f_{\text{loaded}} - f_0$), Δm is the added mass per unit area, and S_m is the sensitivity factor. The numerical value of S_m depends on the design, material, and operating

frequency (wavelength) of the acoustic sensor. Therefore, the oscillating sensor converts the mass value into a frequency shift. Because frequency and time are the easiest variables to measure by electronic circuits, the entire sensor's accuracy is determined virtually by the ability to assure that the coefficient S_m is known and does not change during the measurement (see Fig. 17.18 as an example of this type of a sensor).

Molecules or larger particles of a chemical compound deposit on the surface of the crystal increasing its mass and, subsequently, lowering its resonant frequency. An electronic circuit measures the frequency shift, which is almost a linear measure of the chemical concentration in the sampled gas. Thus, this method is sometimes called a *microgravimetric* technique, as added mass is extremely small. The absolute accuracy of the method depends on such factors as the mechanical clamping of the crystal, temperature, and so forth; therefore, the over-the-range calibration is usually required.

Oscillating sensors are extremely sensitive. For instance, a typical sensitivity is in the range of 5 MHz cm^2/kg, which means that 1 Hz in frequency shift corresponds to about 17 ng/cm^2 added weight. The dynamic range is quite broad: up to 20 μg/cm^2. To assure a selectivity, a crystal is coated with a chemical layer specific for the material of interest.

Another type of a gravimetric detector is a surface-acoustic-wave (SAW) sensor. The SAW is a phenomenon of propagating mechanical waves along a solid surface which is in a contact with a medium of lower density, such as air [21]. These waves are sometimes called Reyleigh waves, after the man who predicted them in 1885. As with a flextural plate, the SAW sensor is a transmission line with three essential components: the piezoelectric transmitter, the transmission line with a chemically selective layer, and the piezoelectric receiver. An electrical oscillator causes the electrodes of the transmitter to flex the substrate, thus producing a mechanical wave. The wave propagates along the transmission surface toward the receiver. The substrate may be fabricated of $LiNbO_3$ with a high piezoelectric coefficient [22]. However, the transmission line does not have to be piezoelectric, which opens several possibilities of designing the sensor of different materials, like silicon. The transmission surface interacts with the sample according to the selectivity of the coating, thus modulating the propagating waves. The waves are received at the other end and converted back to an electric form. Often, there is another reference sensor whose signal is subtracted from the test sensor's output.

Typical designs of the acoustic sensors which can be adapted for measuring mass are covered in Section 12.6 of Chapter 12. Here, we briefly describe the gravimetric SAW sensor which is adapted for sensing gas concentrations (Fig. 17.14). The sensor is designed in the form of a flextural thin silicon plate with two pairs of the interdigitized electrodes deposited by use of the sputtering technology. A thin piezoelectric ZnO thin film is deposited beneath the electrodes, so that the plate can be mechanically excited by the external electronic circuit. The piezoelectric film is needed to give piezoelectric properties to the silicon substrate. The top surface of the sensing plate is coated with a thin layer of a chemically selective material (or glue, if the sensor is intended to detect air pollutants). The entire sensor is positioned inside a tube where the sampled gas is blown through. The left and right pairs of the electrodes are

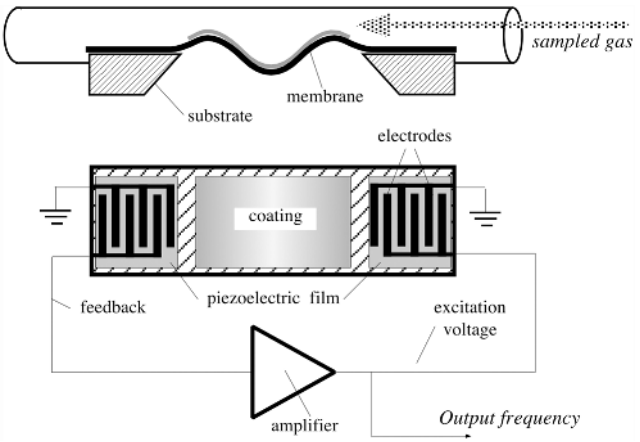


Fig. 17.14. Flextural-plate SAW gas sensor; deflection of the membrane is exaggerated for clarity.

connected to the oscillating circuit whose frequency f_0 is determined by the natural mechanical frequency of the sensor's plate.

The circuit contains an amplifier whose output drives the excitation electrode. Due to the piezoelectric effect, this results in flexing the membrane and propagation of the deflection wave from right to left. The wave velocity is determined by the state of the membrane and its coating. The change in the mechanical properties of the coating depends on its interaction with the sampled gas. Thus, the left electrodes will detect piezoelectric response either sooner or later, depending on how fast the wave goes through the membrane. The received signal is applied to the amplifier's input as a feedback voltage and causes the circuit to oscillate. The output frequency is a measure of the sampled gas concentration. The reference frequency is usually determined before sampling the gas.

One of the possible applications for the technique is the monitoring of heterogeneous samples, such as aerosols and suspensions. The mass increase due to impacting and sticking particles (liquid–aerosol or solid–suspension) produces a strong frequency shift; however, it is also sensitive to particle size, which means that it can be used either to detect the sizes of the particles or to monitor samples with constant particle dimensions. To improve the "stickiness" of the crystal, it can be treated chemically, or an electrostatic effect can be used.

The theoretical sensitivity of the flextural plate sensor is given by $S_m = -1/2\rho\, d$, where ρ is the average density of the plate and d is its thickness [23]. At an operating frequency of 2.6 MHz, the sensor has sensitivity on the order of -900 cm^2/g. So, for example, if the sensor having the area of 0.2 cm^2 captures 10 ng (10^{-8} g) of material, the oscillating frequency is shifted by $\Delta f = -(900)(2.6 \times 10^6)(10^{-8}/0.2) = -117$ Hz.

The SAW sensors are quite versatile and can be adapted for measuring a variety of chemical compounds. The key to their efficiency is the selection of the coating. Table 17.1 gives examples of various SAW sensors.

Table 17.1. SAW Chemical Sensors

Compound	Chemical Coating	SAW Substrate
Organic vapor	Polymer film	Quartz
SO_2	TEA[a]	Lithium niobate
H_2	Pd	Lithium niobate, silicon
NH_3	Pt	Quartz
H_2S	WO_3	Lithium niobate
Water vapor	Hygroscopic	Lithium niobate
NO_2	PC[b]	Lithium niobate, quartz
NO_2, NH_3, NH_3, SO_2, CH_4	PC[b]	Lithium niobate
Vapor explosives, drugs	Polymer	Quartz
SO_2, methane	C[c]	Lithium niobate

Source: Ref. [22].

[a] Triethanolamine.

[b] Phthalocyanine.

[c] No chemical coating used. Detection is based on changes in thermal conductivity produced by the gas.

17.5.5 Biochemical Sensors

Biosensors are a special class of chemical sensors. The evolution of species by means of natural selection led to extremely sensitive organs, which can respond to presence of just few molecules. Man-made sensors, although generally not as sensitive, employ biologically active materials in combination with several physical sensing elements (e.g., amperometric or thermal). The biorecognition element is actually a bioreactor on the top of the conventional sensor, so the response of the biosensor will be determined by the diffusion of the analyte, reaction products, coreactants or interfering species, and the kinetics of the recognition process. The following biological elements may be detected qualitatively and quantitatively by the biosensors: organisms, tissues, cells, organelles, membranes, enzymes, receptors, antibodies, and nucleic acids [17].

In the fabrication of biosensor, one of the key issues is *immobilization* of analytes on the physical transducer. The immobilization must confine the biologically active material on a sensing element and keep it from leaking out over the lifetime of the biosensor, allow contact to the analyte solution, allow any product to diffuse out of the immobilization layer, and not denature the biologically active material. Most of the biologically active materials used in biosensors are proteins or contain proteins in their chemical structures. Therefore, to immobilize the proteins on the surface of the sensor, two basic techniques are employed: binding or physical retention. Adsorption and covalent binding are the two types of binding technique. The retention involves separating the biologically active material from analyte solution with a layer on the surface of the sensor, which is permeable to the analyte and any products of the recognition reaction, but not to the biologically active material.

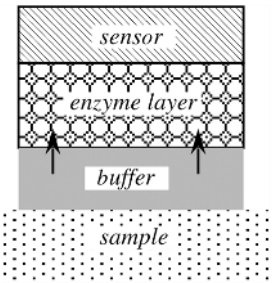


Fig. 17.15. Schematic diagram of an enzyme sensor.

17.5.6 Enzyme Sensors

One of the most efficient ways of achieving selectivity is by using sensors with enzymatic layers. Enzymes are a special kind of catalyst—proteins of molecular weight 6–4000 kDa found in living organisms. They have two remarkable properties: (1) They are extremely selective to a given substrate and (2) they are extraordinarily effective in increasing the rate of reactions. Therefore, they favorably contribute to both the selectivity and the magnitude of the output signal. The maximum velocity of the reaction is proportional to the concentration of the enzyme. A general diagram of an enzymatic sensor is shown in Fig. 17.15 [17].

The sensing element can be a heated probe, an electrochemical sensor, or an optical sensor. Enzymes operate only in an aqueous environment, so they are incorporated into immobilization matrices which are gels—specifically, hydrogels. The basic operating principle is as follows. An enzyme (a catalyst) is immobilized inside a layer into which the substrate diffuses. Hence, it reacts with the substrate and the product is diffused out of the layer into the sample solution. Any other species which participates in the reaction must also diffuse in and out of the layer.

17.6 Chemical Sensors Versus Instruments

Because of the complexity of operation and numerous influences, a chemical sensor is rarely used alone, but, rather, it is a key part of a more rigorous chemical detection instrument. An instrument often combines sensor measurement hardware with decision-making and control software (instructions). Most instruments employ some form of feedback to adjust the operation based on actual conditions versus desired conditions. Some instruments and microinstruments include components to perform mechanical actions (e.g., pumping, filtration, and separation).

Instruments such as gas chromatographs, mass spectrometers, IR spectrometers, and others provide the most comprehensive chemical analysis, especially when compared to simple individual gas sensors. These instruments contain sensors calibrated to perform a specific type of measurement or analysis as well as support circuits and signal processing to control, minimize, and compensate for chemical signal drift and other operation-induced errors.

Liquid and gas chromatography (LC and GC, respectively) are effective and popular chemical analysis methods. Chromatography involves injecting a liquid or gas

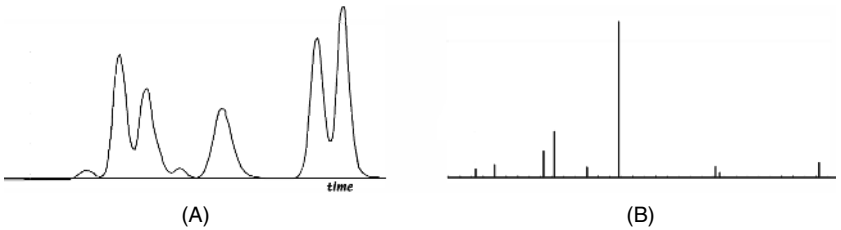


Fig. 17.16. Gas-chromatography example (A) and mass-spectrometry example (B).

analyte into a restrictive tube that is filled with a highly porous material presenting a highly tortuous path to the molecules in the analyte [24]. The size of the pores in the material are chosen by prior experiments to closely match the physical size of the expected molecules in the sample. The adsorption behavior of the material is also matched to the anticipated molecules. An electrical (e.g., conductance) detector resides at the end of the tubing path and recognizes the presence of any molecules that exit the tube. A timer is started when the sample is injected. As the sample passes through the porous material, the smaller molecules move easier and exit sooner than larger molecules. This effectively separates the sample, like passing gravel through a series of sieves. The samples are separated and eventually exit the tube, typically in groups with gaps in between. The electrical detector records their overall concentration as a peak with a width, and an integrated area that is a function of that molecule's concentration in the sample. The time at which each peak is recorded is a function of the molecule size and adsorption characteristics and is used to differentiate molecules and identify them. The resulting peak versus time data are called a "chromatogram" (Fig. 17.16A).

Modern advanced chromatographic systems produce multiple chromatograms and store specialized libraries of samples to allow comparison. Some systems even generate multidimensional chromatograms for special purposes. Chromatography is a popular chemical analysis method with a wide variety of instrument manufacturers and excellent training textbooks available. Chemometrics application software can process standardized chromatography responses using a variety of calibration and pattern recognition techniques.

Mass spectrometry (MS or "mass spec") is a chemical analysis method that involves ionizing the analyte sample and then accelerating the produced ions with a potential and focusing them into a beam [25]. The beam is composed of molecular fragments with different masses and different net charges, each of which is separated into a spectrum with magnetic or electrostatic forces. The result is a mass spectrogram (17.16B). The location of the lines in the spectrogram indicates the mass/charge ratio and is indicative of the molecular fragments (chemical species). The height of the lines is a function of the proportion of molecules of any given m/e ratio in the sample. Mass spec is a popular laboratory method.

Many other chemical analysis techniques are often coupled with mass spec to allow better selectivity for nonionizing compounds. Like chromatography, mass spectrograms can be processed as vectors using Chemometrics techniques and software (see Section 17.6.1).

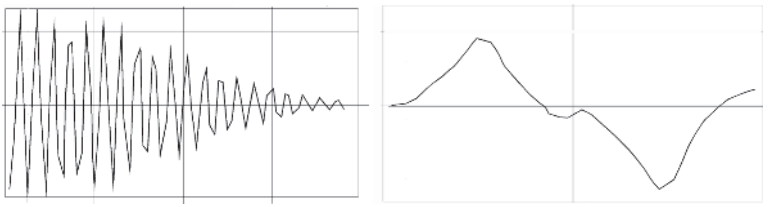


Fig. 17.17. Fourier transform infrared spectroscopy example (A) and voltammetry example (B).

Fourier transform infrared (FTIR) spectroscopy is a chemical analysis technique that involves bombarding an analyte sample with a range of different IR radiations and measuring the magnitude of different wavelengths of the IR that absorbed [26]. The absorption is plotted versus wavelength to produce a very noisy spectrogram that is then filtered via Fourier transforms (Fig. 17.17A). FTIR is still a popular experimental technology with improvements being made regularly, but those improvements also represent changes that make standardization and universal libraries of spectra of limited use. The technique has not yet been miniaturized and the noise found in the spectrum requires significant computational capability to remove it. Chemometrics application software that includes Fourier transforms may be used to process the FTIR spectrograms.

Voltammetry is an electrochemical measurement technique that involves applying a changing potential across two or three electrodes in contact with a liquid or gaseous analyte [27]. The changing potential triggers redox reactions in electroactive species and affects the overall electrical current measurable in the system loop. The plot (Fig. 17.17B) of the measured current versus the applied potential is called a voltammogram [28] and it contains a significant amount of information which allows one to identify and quantify the chemical species in the mixture or compound. The voltammogram produced can be simple or complex depending on the complexity of the shape of the applied potential [29]. There is a great deal of interaction between the different electrochemical reactions, but, in general, different chemical species have specific dissociation potentials, so the location along the potential curve of the feature identifies the species. The size of the features is controlled by the amount of any given species in the analyte [30]. Voltammetric analysis began in the early part of the twentieth century [31,32]. It is an excellent technique for organic, inorganic, metallic, and metallorganic species. Because voltammetry produces such complex results, specialized chemometrics strategies are replacing manual analysis of the voltammogram [33].

Several basic waveforms are effective and popular for producing the voltammetry response. These include a simple linear sweep, a triangle sweep, stair-step sweep, pulsed differential sweep, and a square-wave sweep. The simpler linear sweep and triangle sweep are excellent for diagnostic capability but have somewhat poor detection limits, often limited to 10^{-3}–$10^{-4}\,M$ levels. Stair-step, differential pulse, and square-wave voltammetries have detection limits as low as the 10^{-7}–$10^{-8}\,M$ levels.

17.6.1 Chemometrics

Often, chemical analysis instruments produce a response signal that is complex and contains a significant amount of information. This requires more advanced analysis than simple single-value measurements or threshold comparisons. It is important to relate the required signal processing to the chemical sensor, as they influence each other quite significantly. The study of chemical measurements is referred to as *chemometrics* and has been developed to address the particular challenges of chemical sensors and complex chemical response analysis. Chemometrics is responsible for a great deal of advanced data processing techniques, applying mathematical and statistical modeling to chemical systems [34]. In general, these techniques can be divided into *data exploration* and data analysis topics.

Part of both the exploration and analysis involves *modeling* the response data. Models can be divided into *parametric* and *nonparametric* types [35]. Statistical techniques were once taught based on strong assumption about the data—assumptions that continuous variables followed Gaussian (Normal) distributions. This also describes any fitting operation where a preassumed shape is assigned and the data fit against it. The error is measured as deviation from the assumed fit. Statistical methods that make strict assumptions about the distribution for experimental data are referred to as *parametric methods* and produce exact solutions to approximate problems.

Approximate solutions to exact problems, the complement, make no assumptions about distributions. These *nonparametric models* are usually easier and quicker to apply, with a simple theory that allows better judgment to be used in their application. *Robust statistical models* are an alternative to strictly parametric or nonparametric methods. Robust statistics attempts to describe the structure best fitting the *bulk* of the data and to identify outliers and leverage points (influential points that have large affect on regressions) for optional separate treatment. Robust statistics also deals with unsuspected serial correlations or deviations from assumed serial correlations.

Data exploration techniques typically start with *unsupervised classification*. Unsupervised classification (clustering) is a good way to identify and display *natural* grouping without imposing any prior class membership. *Hierarchical cluster analysis* (HCA) is a popular way to implement unsupervised classification. HCA is implemented by calculating all point-to-point distances, sorting them, and then, starting with smallest distances, linking points together to form new seed clusters. Points or clusters are joined to the their nearest neighbor (based on Euclidean distance), forming a growing chain of links until the entire population is assigned.

Data analysis also includes classification methods, but these *supervised classification* techniques are used to construct a model to classify future samples. Many options exist for performing supervised classification, but all the methods share one common trait; They use past and existing examples of classified response to assign unknown responses to various groupings or categories. This is the *supervised* aspect to the classification. In the K-nearest neighbors algorithm (KNN), an unknown sample is assigned to the class that it is nearest to in multidimensional (Euclidean) space [34]. Another supervised classification approach that also reduces the number of variables/dimensions is soft independent modeling of class analogy (SIMCA) [36].

SIMCA performs better with lower sample: variable ratios than other supervised classification methods.

17.6.2 Multisensor Arrays

Processing multiple measurements from individual chemical sensors and from a number of independent sensors can provide information needed to statistically reduce error and improve both the selectivity and sensitivity of a chemical sensor [37] or chemical detection instrument. Because measurement error is a sum of systematic error and random error, the measurement error of an individual sensor can be statistically reduced via multiple samples by using statistics to reduce or eliminate the random error [36]. Multiple redundant sampling can provide enough data to reduce the measurement standard deviation by a factor of $1/\sqrt{n}$, where n is the number of redundant samples. The redundant samples may come from the same sensor or multiple sensors of the same type to further ensure the best possible response [38]. This, however, is useful against random errors but is not efficient against systematic errors.

Responses from multiple independent sensors of *different* types can be combined (often referred to as sensor fusion) to provide overlapping reinforced responses that better span the sensors' response spaces, leaving fewer gaps where analyte identification would be weak or unavailable.

Obviously, introducing any redundancy of sensors or multiplicity of measurements increases the amount of data and the complexity of signal processing. There is a trade-off decision to be made between the additional work created and the quality of the decision that can be made with that corresponding data. Often, the majority of improvements can be made to the measurement accuracy with only a limited number of multiple measurements. Significant additional effort typically only gains a small amount of additional accuracy.

17.6.3 Electronic Noses (Olfactory Sensors)

The principle of measurement and data processing that is described in this subsection is an example of a bionic approach of resolving the selectivity and sensitivity. The main idea is to use many sensors of different types and process data in a way that resembles data processing by living brains. Although today we still know very little of how brain really works, some ideas suggested by Nature already can be put to a practical use. Processing and analyzing the signals produced by multisensor arrays typically involves pattern recognition. Electronic noses, or *e-noses*, are less a sensor or instrument and more a *measurement strategy*. Electronic noses have become popular and combine advanced sensors and sensor array strategies with chemometrics techniques to produce a broad range of intermediate instruments and analyzers.

Early e-noses tried to duplicate the behavior and capability of human odor sensing. They combined different sensor types to represent the different cell tissues in the nasal cavity and they took the approach of detecting an odor as a collection of individual chemicals. The name "odor sensor" is used instead of "gas sensor" whenever its sensitivity approaches that of a human. Odor and fragrance sensors find applications

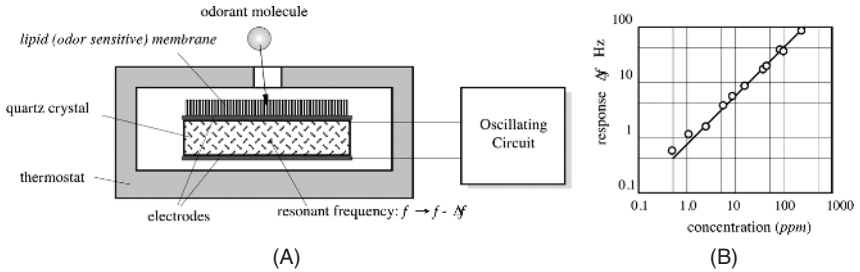


Fig. 17.18. Microbalance odor sensor (A) and its transfer function (B) for amylacetate gas.

in forensic science, quality assurance in the cosmetic and food industry, environmental control, and so forth. All methods of odor measurements can be divided into four groups: instrumentational analysis, semiconductor gas sensors, membrane potential-type odor sensors [39], and the quartz microbalance method. The last method is conceptually close to the gravimetric sensors covered in previous sections. In general, it is based on a shift in natural frequency of a quartz crystal coated with an odor-sensitive membrane and the subsequent measurement of the shift (Fig. 17.18). This can be measured by electronic means and correlated with the odorant concentration. Potentially, this method has the possibility of performing with humanlike characteristics and sensitivity because the same membrane as the human olfactory (lipid membrane) can be used as the odorant-absorptive media of the sensor.

Olfactory cells or odor receptors of humans are covered with a phospholipid bilayer membrane, which is a kind of lipid membrane. It is believed that odorant molecule adsorption into the membrane induces nerve pulses. Using this as an analogy, a man-made odor sensor uses a composite membrane consisting of PVC, a plastisizer, and synthetic lipid [40]. The synthetic lipid molecules are randomly oriented in the polymer matrix. To produce a sensor, a quartz crystal was cut to 14 mm in diameter. Then, the lipid composite was prepared as a solution of organic solvent (tetrahydrofuran), PVC, plastisizer (dioctylphenyl phosphonate), and synthetic lipid (dioctil phosphate, decyl alcohol, and other lipids can be employed). The membrane is formed with a thickness of 200 μm on one side of the resonator by using the spin-coating method (see Chapter 18). The membrane blend is selected to maintain the quality factor of the resonator (Q) on the level of at least 5×10^4.

The experimental curve of the transfer function indicates that the response was detectable starting from 1 ppm concentration, which is approximately equal to the human threshold, and was linear up to a concentration of about 3000 ppm. Such a sensor has a quite fast response time—within 1 s.

Newer approaches to e-nose development involve more flexible combinations of sensor designs and signal processing. The performance of these e-noses is measured more by how many compounds they can distinguish at nominal low ppm levels and less by their sensitivity and detection limit for a specific compound. Because most chemical sensors are affected by both humidity and temperature, sensors for such conditions are often included in the e-nose array [41]. One example of an experimental

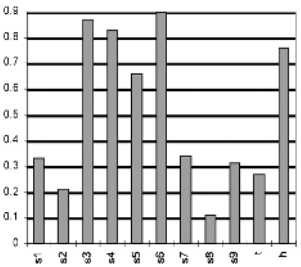


Fig. 17.19. Metal-oxide e-nose response array.

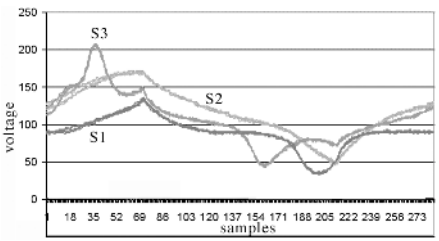


Fig. 17.20. Background air readings from each of the three sensors.

e-nose employed a set of nine simple, but specialized, commercial tin-dioxide gas sensors. Each metal-oxide device was doped, making the metal oxide more specific to a particular gas species. The simple time-based conductivity change responses from the devices were collected into an array response, as shown in Fig. 17.19.

The combination of devices could differentiate common office chemicals such as contact cement, paint thinner, glass cleaner, and alcohol. Using these collections of sensors experiments achieved up to a 98% accuracy identification rate.

Another example electronic nose developed for fire detection employed a complementary strategy: combining fewer but more complex sensors in a smaller array [37]. The signatures from three electrochemical sensors were very different, as shown in Fig. 17.20, and were fused to produce a complex temporal array signature.

These solid electrochemical sensor arrays were used to characterize various combustible materials commonly found on naval ships such as wood, wallboard, cleaning fluid, plastics, food, bedding materials, and fabrication shop operations such as welding. The approach employed a time-offset signature series, where the *change* in signatures was monitored over time as opposed to the *static* signature. Using this approach, the miniature-array e-nose was able to correctly identify 14 different types of fire with a confidence of between 70% and 100%.

The combination of different types of chemical sensor (from only a couple to tens of devices) allows overlaps in their respective detection ranges to complement each other, producing higher-quality detection from simpler, less selective sensors than any individual sensor could achieve on its own. A generalized but also complete and functional model for an e-nose can be constructed that includes both simple and complex chemical sensors, along with ancillary sensors such as temperature, humidity, and barometric pressure to measure the effects of these variables on the chemistry. For most problems, the output will be mapped to specific categories, including an

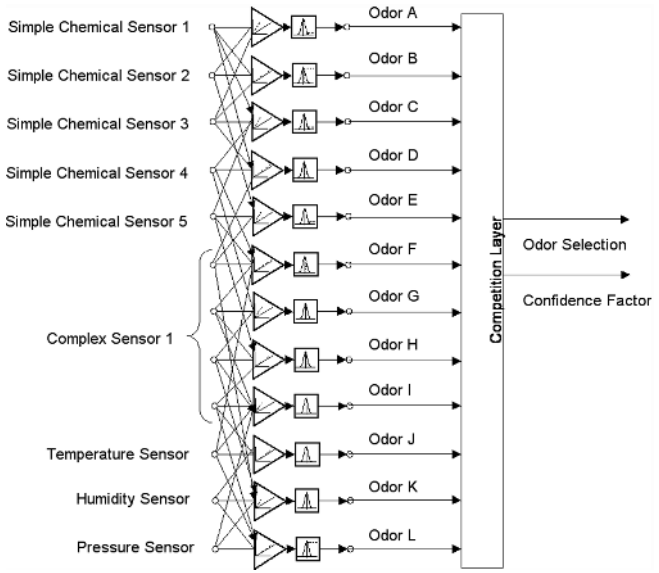


Fig. 17.21. Generalized e-nose model (for simplicity, only a portion of the sensor connections are represented; the full model would include connections from *all* sensors to *all* categories).

"unknown" category for samples which are detected, but do not fall into existing categories within some predetermined level of confidence. Such a model is shown in Fig. 17.21.

This detection/sensing model is particularly appropriate for identifying complex mixtures and ratios of chemical constituents as a group, rather than isolating and quantifying any particular single-gas species. Because of this fundamental underlying strategy, e-noses are particularly popular in food industry and process control, where they are used to categorize beverages, grade the quality of extracts, and even determine the age and expiration dates of produce. These are tasks that historically are very subjective and qualitative when performed by a human expert, but become far more reproducible when performed by an e-nose.

17.6.4 Neural Network Signal (Signature) Processing for Electronic Noses

Active array devices like an e-nose produce complex signals or "signatures," which have to be processed to extract the desired chemical species component information. It is natural and effective to pair e-nose signals with neural network classification and analysis methods that similarly mimic biological systems [37].

Neural network algorithms can duplicate the more preferred chemometrics pattern recognition methods, such as Bayesian classifiers, providing provable and statistically measurable confidence in their results. Neural methods execute simple mathematical operations in a highly parallel fashion and lend themselves to scalable execution from low-cost microcontrollers.

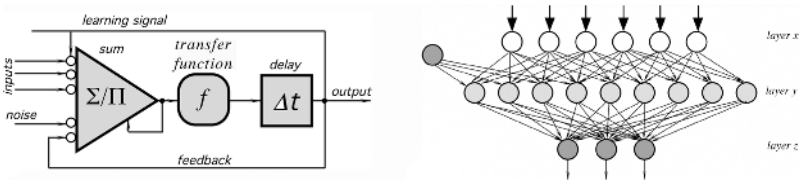


Fig. 17.22. Generalized neuron model (A) and layers combined into a network (B).

A neural network is inspired by and loosely models the architecture and information processing capability of the biological brain [42]. An artificial neural network (ANN) accomplishes this by simulating each biological neuron with an integrated circuit as a collection of gates and transistors, whereas a computational neural network (CNN) accomplishes this through execution of a series of computer instructions. Neural networks can be structured to perform classification [43], to approximate equations [44], and to predict values [45,46]. Several different models for neurons are available; each supports a different range of network architectures and artificial learning methods. A generalized neuron model (Fig. 17.22A) includes some input stage with variable weighted interconnections to the outputs of other neurons, a summation/comparison stage for combining the weighted inputs, a transfer function that reduces the information passed along through the neuron, an output stage that connects to the inputs of other neurons, and some feedback/training method to adjust the weights so that a desired output is produced when exposed to known inputs. Some network architectures require an optional delay stage to support adaptive learning.

A generalized network architecture (Fig. 17.22B) includes an input layer x that interfaces directly to the sensor signals, a hidden layer y that reduces information, makes intermediate choices, and performs feature extraction, and an output layer z that selects intermediate answers and provides the classification or component analysis information. In a generic architecture, neurons are referred to as nodes, and internode connections are only made between adjacent layers.

Electronic noses generally pursue composite odor *classification*, with component *analysis* representing a more difficult secondary goal. Probabilistic neural network (PNN) classifiers are the most popular CNNs used with electronic noses. They duplicate the functionality of K-nearest neighbor or Bayesian statistical classifiers, though the NN versions often outperform both [47]. The PNN uses a radial basis function neuron and competitive hidden layer network architecture. PNNs require supervised training where a set of inputs is constructed that has predetermined desired outputs (categories). During training, a new neuron is constructed for each sample in the training set. The weights between the inputs and the competitive neuron are copies of the input values themselves. The output of each neuron goes to a matching category in the final competitive output layer. Multiple examples of a given input/output pairing create additional copies of a neuron and strengthen the possibility of selection

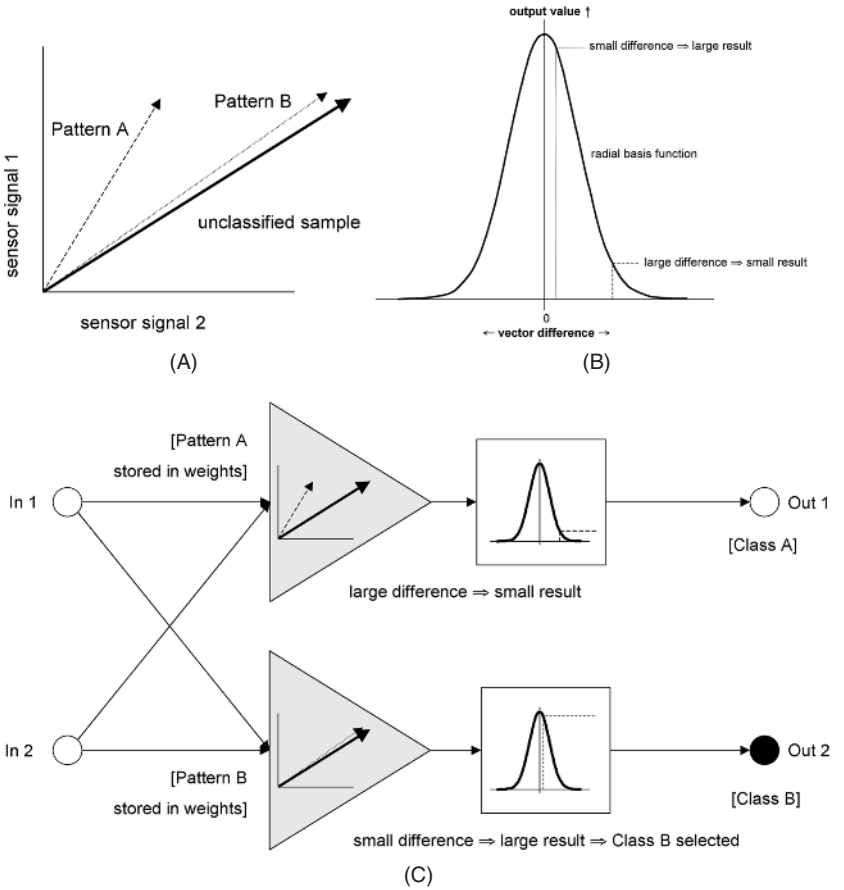


Fig. 17.23. Vector comparison (A), radial basis function (B), and PNN layers (C).

for that category, reflecting statistical probabilities of that category's occurrence in a population—hence, the name "probabilistic" neural network.

During the operation of a PNN, a vector containing the input values is presented to each neuron in the input layer. Each neuron compares the input vector and the vector formed from its own local set of weights by computing a Cartesian distance between the vectors (Fig. 17.23A). Internal to each neuron, the distance is then passed through the local radial basis transfer function (a Gaussian bell curve centered on input = 0 to produce an output = 1) that outputs a high value for small distances (differences) and very small values for larger distances (Fig. 17.23B). The result is that the neuron whose weights most closely match the input vector produces the highest final output value, and the output layer assigns the input to that category (Fig. 17.23C). The options for training PNNs vary with trade-offs among flexibility, memory resource use, and speed of training.

17.6.5 "Smart" Chemical Sensors

Many other chemical sensors, both commercial and experimental, employ a growing variety of phenomena and strategies. Trends in microelectronics and programmable controllers will lead to the production of "smart" chemical sensors. The future of chemical sensors lies in these smart devices. A smart sensor incorporates some level of the data processing into the sensor directly, distributing some of the intelligence of the instrument and allowing functional and useful systems to be designed with lower intelligence required in the instrument [48–50]. A smart *chemical* sensor should include interdevice communication and local drift and recalibration capabilities so that remote polling control systems would only receive measurements. A smart chemical sensor may also perform routine unit conversion (i.e., from % to ppm) and report different units to different requests. In this way, the same (smart) sensor can provide a measurement to different hosts without requiring any of them to introduce any additional scaling of their own; they work with whatever local units they chose.

References

1. Edmonds, T.E. (ed). *Chemical Sensors*. Blackie and Son, New York, 1988.
2. *General Information for TGS Sensors, Rev. 6.98*. Figaro USA Inc., Glenview, IL, 1998.
3. Sberveglieri, G. (ed.). *Gas Sensors: Principles, Operations, and Developments*, Kluwer Academic, Boston, MA, 1992, pp. 8, 148, 282, 346–408.
4. Blum, L.J., *Bio- and Chemi-Luminescent Sensors*, World Scientific, River Edge, NJ, 1997, pp. 6–32.
5. Smith, J.A, Polk B.J., Kikas, T., and Levermore, D.M. ChemFETs: Chemical sensors for the real world. www.bizoki.chemistry.gatech.edu/janata-chemical-sensors, 2000.
6. Wróblewski, W., Dawgul, M., Torbicz, W., and Brzózka, Z. Anion-selective CHEMFETs, Department of Analytical Chemistry, Warsaw University of Technology, Warsaw, 2000.
7. Hydrogen sensor (white paper), Sandia National Laboratory, Sandia, NM, 2002; abailable from www.sandia.gov/mstc/technologies/microsensors/techinfo.
8. Gentry, S. J. Catalytic devices. In: *Chemical Sensors*. Edmonds, T. E. (ed.). Chapman & Hall, New York, 1988.
9. Cobbold, R.S.C. *Transducers for Biomedical Measurements*. John Wiley & Sons, New York, 1974.
10. Tan, T.C. and Liu, C.C. Principles and fabrication materials of electrochemical sensors. In: *Chemical Sensor Technology*. Kodansha Ltd., 1991, Vol. 3.
11. Clark, L.C. Monitor and control of blood and tissue oxygen tension. *Trans. Am. Soc. Artif. Internal Org.* 2, 41–46, 1956.
12. Vogt, M. C., Shoemaker, E. L., MacShane, D. A., and Turner, T. An intelligent gas microsensor employing neural network technology. *J. Appl. Sensing Technol.* September, 54–62, 1996.

13. LaCourse, W.R. *Pulsed Electrochemical Detection in High-Performance Liquid Chromatography*, John Wiley & Sons, New York, 1997, pp. 13–20, 49, 136, 173, 258–259.
14. Skubal, L.R., Meshkov, N.K., and Vogt, M.C. Detection and identification of gaseous organics using a TiO_2 sensor, *J. Photochem. Photobiol. A: Chem.*, 148, 103–108, 2002.
15. Severin, E. *Cyrano Sciences' Sensor Technology—The heart of the Cyranose 320 Electronic Nose*. Cyrano Sciences Inc., 2000; www.cyranosciences.com/technology/sensor.
16. Hydrocarbon fuel, HCl sensor look for trouble. *Sensors*, 11–12, 1991.
17. Dewa, A.S. and Ko, W.H. Biosensors. In: *Semiconductor Sensors*. Sze, S.M, (ed.). John Wiley & Sons, New York, 1994, pp. 415–472.
18. Morgan, C.H. and Cheung, P.W. An integrated optoelectronic CO_2 gas sensor. In: *Transducers'91. International Conference on Solid-State Sensors and Actuators. Digest of Technical Papers*. IEEE, New York, 1991, pp. 343–346.
19. Dybko, A. and Wroblewski, W. Fiber optic chemical sensors. www.ch.pw.edu.pl/∼dybko/csrg/fiber/operating, 2000.
20. Seiler, K. and Simon, W. Principles and mechanisms of ion-selective optodes. *Sensors Actuators B* 6, 295–298, 1992.
21. Ristic, V.M., *Principles of Acoustic Devices*. John Wiley & Sons, New York, 1983.
22. Nieuwenhuizen, M.S., et al. Transduction mechanism in SAW gas sensors. *Electron. Lett.* 22, 184–185, 1986.
23. Wenzel, S.W. and While, R.M. Analytic comparison of the sensitivities of bulk-surface-, and flextural plate-mode ultrasonic gravimetric sensors. *Appl. Phys. Lett.*, 54, 1976–1978, 1989.
24. Malmstadt, H.V., Enke, C.G., Crouch, S.R., and Horlick, G. *Electronic Measurements for Scientists*. W. A. Benjamin, Menlo Park, CA, 1974.
25. Wade, L.G. *Organic Chemistry*, Prentice-Hall, Englewood Cliff, NJ, 1987.
26. Smith, B.C. *Fundamentals of Fourier Transform Infrared Spectroscopy*. CRC Press, New York, 1995.
27. Smyth, M. R. and Vos, J. G. *Comprehensive Analytical Chemistry–Analytical Voltammetry*. Elsevier Science, New York, 1992, Vol. 27, pp. 20, 34, 59.
28. Bard, A.J. and Faulkner, L.R. *Electrochemical Methods*, John Wiley & Sons, New York, 1980, pp. 232–236.
29. Kumta, P.N., Manthiram, A., Sundaram, S.K. and Chiang, Y.M. (eds.). *Processing and Characterization of Electrochemical Materials and Devices*. American Ceramic Society, Westerville, OH, 2000, p. 379.
30. Albery, W.J. and Haggett, B.G.D. New electroanalytical techniques. Electrochemical detectors—fundamental aspects and analytical applications. Proceedings of a Symposium Sponsored by the Analytical and Faraday Division of the Royal Society of Chemistry, Ryan, T.H. (ed.). 1984, p. 15.
31. Scholander, A. *Introduction to Practical Polarography*. Jul. Gjellerups Forlag, Radiometer, Copenhagen, 1950.
32. Heyrovsky, J. and Zuman, P. *Practical Polarography. An Introduction for Chemistry Students*. Academic Press, New York, 1968.

33. *Handbook of Electroanalytical Products*, Bioanalytical Systems Inc., West Lafayette, IN, 1997.
34. Beebe, K.R., Pell, R.J. and Seasholtz, M.B. *Chemometrics. A Practical Guide*. John Wiley & Sons, New York, 1998.
35. Haswell, S.J. (ed.). *Practical Guide to Chemometrics*. Marcel Dekker, New York, 1992, pp. 39–43, 225–226, 310.
36. Einax, J.W., Zwanziger, H.W. and Geib, S. *Chemometrics in Environmental Analysis*. VCH, Weinheim, 1997, pp: 2-75.
37. Gottuk, D.T., Hill, S.A., Schemel, C.F. , Strehlen, B.D., Rose-Pehrsson, S.L., Shaffer, R.E., Tatem, P.A., and Williams, F.W. Identification of fire signatures for shipboard multi-criteria fire detection systems. Report No. NRL/MR/6180-99-8386, Naval Research Laboratory, Washington, DC, 1999, pp. 48–87.
38. Prasad, L., Iyengar, S.S., Rao, R.L., and Kashyap, R.L. Fault-tolerant sensor integration using multiresolution decomposition. *Phys. Rev. E.* 49(4B), 3452–3461, 1994.
39. Miyazaki, Y., et al. Responses of monolayer membranes of thiol-containing lipids to odor substances. *Jpn. J. Appl. Phys.*, 31, 1555–1560, 1992.
40. Matsuno, G., et al. A quartz crystal microbalance-type odor sensor using PVC-blended lipid membrane. *IEEE Trans. Instrum. and Meas.* 44(3), 739–742, 1995.
41. Keller, P.E., Kangas, L.J., Liden, L.H., Hashem, S., and Kouzes, R.T. PNNL Document Number: PNL-SA-26597, Pacific Northwest National Laboratory, Richland, WA, 1996.
42. Masters, T. *Practical Neural Network Recipes in C++*. Academic Press, Boston, MA, 1993, pp. 174–185.
43. Raimundo, I.M. and Narayanaswamy, R. Simultaneous determination of relative humidity and ammonia in air employing an optical fiber sensor and artificial neural network. *Sensors Actuators B: Chem.* 74(1–3), 60–68, 2001.
44. Joo, B.S., Choi, N.J., Lee, Y.S., Lim, J.W., Kang, B.H., and Lee, D.D. Pattern recognition of gas sensor array using characteristics of impedance. *Sensors Actuators B: Chem.*, 77(1–2), 209–214, 2001.
45. Freeman, J. and Skapura, D. *Neural Networks, Algorithms, Applications, and Programming Techniques*. Addison-Wesley, Reading, MA, 1991, pp. 89–111.
46. Winquist, F, Hornsten, E.G., Sundgren, H., and Lundstrom, I. Performance of an electronic nose for quality estimation of ground meat. *Meas. Sci. Technol.*, 4(12), 1493–1500, 1993.
47. Stetter, J.R., Findlay, M.W., Schroeder, K.M., Yue, C., and Penrose, W.R. Quality classification of grain using a sensor array and pattern-recognition, *Anal. Chem. Act.*, 284(1), 1–11, 1993.
48. Nwagboso, C.O. (ed.). *Automotive Sensory Systems*. Chapman & Hall, New York, 1993, pp. 324–336.
49. Harsanyi, G. *Sensors in Biomedical Applications Fundamentals. Technology and Applications*. Technomic, Lancaster, PA, 2000, pp. 4–6, 65–67, 191, 295.
50. Kavanagh, R.C. Probabilistic learning technique for improved accuracy of sinusoidal encoders. *IEEE Trans. Ind. Electron.*, 48(3), pp. 673–681, 2001.

18

Sensor Materials and Technologies

Methods of sensor fabrication are numerous and specific for each particular design. They comprise processing of semiconductors, optical components, metals, ceramics, and plastics. Here, we briefly describe some materials and the most often used techniques.

18.1 Materials

18.1.1 Silicon as a Sensing Material

Silicon is present in the Sun and stars and is a principle component of a class of meteorites known as *aerolites*. Silicon is the second most abundant material on Earth, being exceeded only by oxygen; it makes up to 25.7% of the Earth's crust, by weight. Silicon is not found free in nature, but occurs chiefly as the oxide and as silicates. Some oxides are sand, quartz, rock crystal, amethyst, clay, mica, and so forth. Silicon is prepared by heating silica and carbon in an electric furnace, using carbon electrodes. There are also several other methods for preparing the element. Crystalline silicon has a metallic luster and grayish color[1]. The Czochralski process is commonly used to produce single crystals of silicon used for the solid-state semiconductors and micromachined sensors. Silicon is a relatively inert element, but it is attacked by halogens and dilute alkali. Most acids, except hydrofluoric, do not affect it. Elemental silicon transmits infrared radiation and is commonly used as windows in far-infrared sensors.

Silicon's atomic weight is 28.0855, and its atomic number is 14. Its melting point is 1410°C and the boiling point is 23°C. The specific gravity at 25°C is 2.33 and its valence is 4.

Properties of silicon are well studied and its applications to sensor designs have been extensively researched worldwide. The material is inexpensive and can now be produced and processed controllably to unparalleled standards of purity and perfection. Silicon exhibits a number of physical effects which are quite useful for sensor applications (see Table 18.1).

[1] Silicon should not be confused with silicone, which is made by hydrolyzing silicon organic chloride, such as dimethyl silicon chloride. Silicones are used as insulators, lubricants, and for the production of silicone rubber.

Table 18.1. Stimuli of Silicon-Based Sensors

Stimuli	Effects
Radiant	Photovoltaic effect, photoelectric effect, photoconductivity, photo-magneto-electric effect
Mechanical	Piezoresistivity, lateral photoelectric effect, lateral photovoltaic effect
Thermal	Seebeck effect, temperature dependence of conductivity and junction, Nernst effect
Magnetic	Hall effect, magnetoresistance, Suhi effect
Chemical	Ion sensitivity

Source: Ref. [1].

Unfortunately, silicon does not posses the piezoelectric effect. Most effects of silicon such as the Hall effect, the Seebeck effect, piezoresistance, and so forth are quite large; however, a major problem with silicon is that its responses to many stimuli show substantial temperature sensitivity. For instance: strain, light, and magnetic field responses are temperature dependent. When silicon does not display the proper effect, it is possible to deposit layers of materials with the desired sensitivity on top of the silicon substrate. For instance, sputtering of ZnO thin films is used to form piezoelectric transducers which are useful for the fabrication of SAW (surface acoustic waves) devices and accelerometers. In the later case, the strain at the support end of the an etched micromechanical cantilever is detected by a ZnO overlay.

Silicon itself exhibits very useful mechanical properties which currently are widely used to fabricate such devices as pressure transducers, temperature sensors, force and tactile detectors by employing the MEMS technologies. Thin film and photolithographic fabrication procedures make it possible to realize a great variety of extremely small, high-precision mechanical structures using the same processes that have been developed for electronic circuits. High-volume batch-fabrication techniques can be utilized in the manufacture of complex, miniaturized mechanical components which may not be possible with other methods. Table A.14 in the Appendix presents a comparative list of mechanical characteristics of silicon and other popular crystalline materials.

Although single-crystal silicon (SCS) is a brittle material, yielding catastrophically (not unlike most oxide-based glasses) rather than deforming plastically (like most metals), it certainly is not as fragile as is often believed. Young's modulus of silicon (1.9×10^{12} dyn/cm or 27×10^6 psi), for example, has a value of that approaching stainless steel and is well above that of quartz and of most glasses. The misconception that silicon is extremely fragile is based on the fact that it is often obtained in thin slices (5–13-cm-diameter wafers) which are only 250–500 μm thick. Even stainless steel at these dimensions is very easy to deform inelastically.

As mentioned earlier, many of the structural and mechanical disadvantages of SCS can be alleviated by the deposition of thin films. Sputtered quartz, for example, is utilized routinely by industry to passivate integrated circuit chips against airborne impurities and mild atmospheric corrosion effects. Another example is a deposition of silicon nitrate (Table A.14) which has a hardness second only to diamond. Anisotropic

etching is a key technology for the micromachining of miniature three-dimensional structures in silicon. Two etching systems are of practical interest. One is based on ethylenediamine and water with some additives. The other consists of purely inorganic alkaline solutions like KOH, NaOH, or LiOH.

Forming the so-called *polysilicon* (PS) materials allows one to develop sensors with unique characteristics. Polysilicon layers (on the order of 0.5 µm) may be formed by vacuum deposition onto oxided silicon wafer with an oxide thickness of about 0.1 µm [2]. Polysilicon structures are doped with boron by a technique known in the semiconductor industry as LPCVD (low-pressure chemical vapor deposition).

Figure 18.1A shows the resistivity of boron-doped LPCVD polysilicon in a comparison with SCS. The resistivity of PS layers is always higher than that of a single-crystal material, even when the boron concentration is very high. At low doping concentrations, the resistivity climbs rapidly, so that only the impurity concentration range is of interest to a sensor fabrication. The resistance change of PS with temperature is not linear. The temperature coefficient of resistance may be selected over a wide range, both positive and negative, through selected doping (Fig. 18.1B). Generally, the temperature coefficient of resistance increases with decreased doping concentration. The resistance at any given temperature of a PS layer may be found from

$$R(T) = R_{20} e^{\alpha_R (T - T_0)}, \tag{18.1}$$

where

$$\alpha_R = \frac{1}{R_{20}} \frac{dR(T_0)}{dT}$$

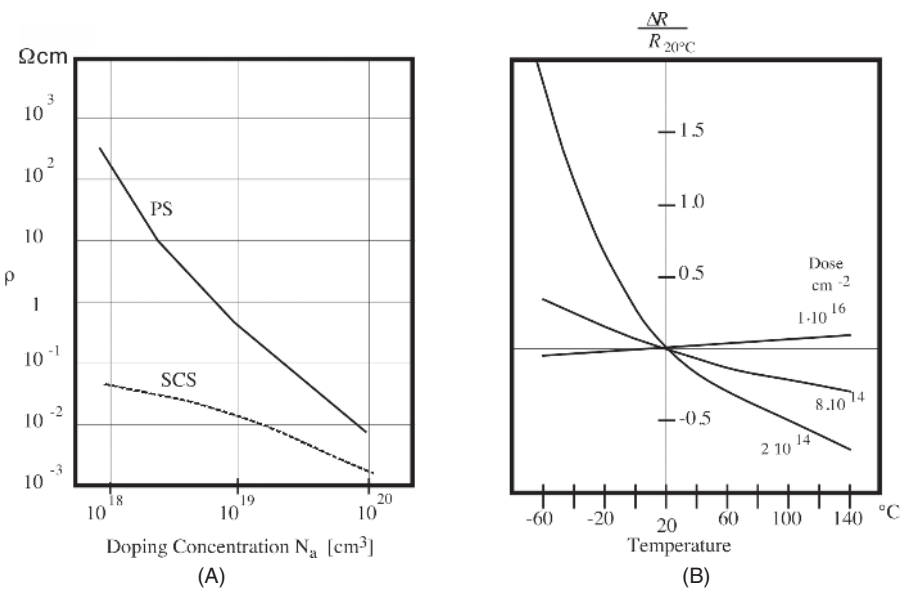


Fig. 18.1. Specific resistivity of boron-doped silicon (A); temperature coefficient of resistivity of silicon for different doping concentrations (B).

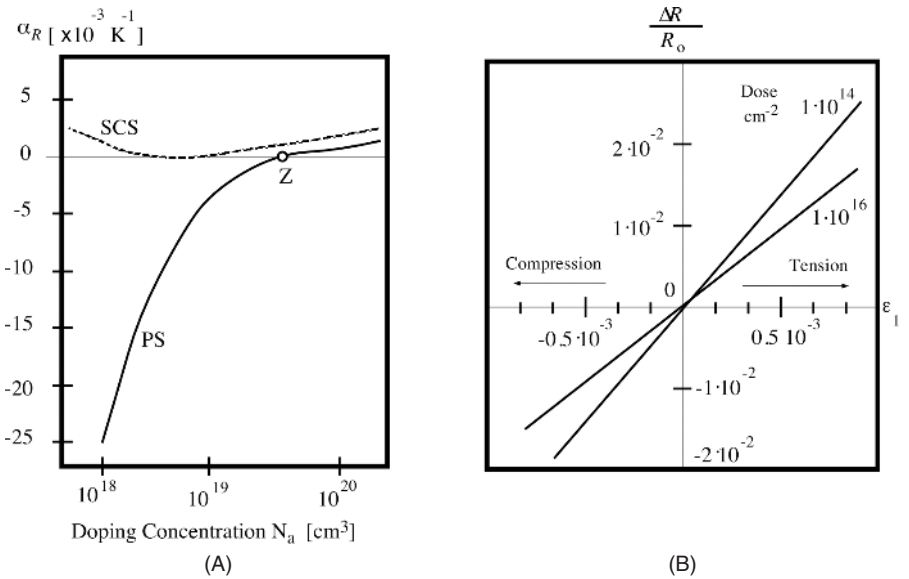


Fig. 18.2. Temperature coefficient as function of doping (A) and piezoresistive sensitivity of silicon (B).

is the temperature coefficient and R_{20} is the resistance at the calibrating point ($T_0 =$ 20°C). Figure 18.2A shows that the temperature sensitivity of PS is substantially higher than that of SCS and can be controlled by doping. It is interesting to note that at a specific doping concentration, the resistance becomes insensitive to temperature variations (point Z).

For the development of sensors for pressure, force, or acceleration, it is critical to know the strain sensitivity of PS resistors expressed through the gauge factor. Figure 18.2B shows curves of the relative resistance change of boron-doped PS resistors, referenced to the resistance value R_0 under no-stress conditions, as a function of longitudinal strain ε_1. The parameter varies with the implantation dose. It can be seen that the resistance decreases with compression and increases under tension. It should be noted that the gauge factor (a slope of the line in Fig. 18.2B) is temperature dependent. PS resistors are capable of realizing at least as high a level of long-term stability as any that can be expected from resistors in SCS, because surface effects play only a secondary role in device characteristics.

18.1.2 Plastics

Plastics are synthetic materials made from chemical raw materials called monomers. A monomer (one chemical unit) such as ethylene is reacted with other monomer molecules to form long chains of repeating ethylene units, forming the polymer polyethylene. In a similar manner, polystyrene is formed from styrene monomers. The polymers consist of carbon atoms in combination with other elements. Polymer

Element	Atomic weight	Energy Bonds	
Hydrogen	1	−H	1
Carbon	12		4
Nitrogen	14		3
Oxygen	16	−O−	2
Fluorine	19	−F	1
Silicon	28		4
Sulfur	32	−S−	2
Chlorine	35	−Cl	1

Fig. 18.3. The atomic building blocks for polymers.

chemists use only eight elements to create thousands of different plastics. These elements are carbon (C), hydrogen (H), nitrogen (N), oxygen (O), fluorine (F), silicon (Si), sulfur (S), and chlorine (Cl). Combining these elements in various ways produces extremely large and complex molecules.

Each atom has a limited capacity (energy bonds) for joining with other atoms, and every atom within a molecule must have all of its energy bonds satisfied if the compound is to be stable. For example, hydrogen can bond only to one other atom, whereas carbon or silicon must attach to four other atoms to satisfy its energy bonds. Thus, H-H and H-F are stable molecules, whereas C-H and Si-Cl are not. Figure 18.3 shows all eight atoms and the corresponding energy bonds.

Adding more carbon atoms in a chain and more hydrogen atoms to each carbon atom creates heavier molecules. For example, ethane gas (C_2H_6) is heavier than methane gas because it contains additional carbon and two hydrogen atoms. Its molecular weight is 30. Then, the molecular weight can be increased in increments of 14 (1 carbon + 2 hydrogen), until the compound pentane (C_5H_{12}) is reached. It is too heavy to be gas and, indeed, it is liquid at room temperature. Further additions of CH_2 groups makes progressively a heavier liquid until $C_{18}H_{38}$ is reached. It is solid: paraffin wax. If we continue and grow larger molecules, the wax becomes harder and harder. At about $C_{100}H_{202}$, the material with a molecular weight of 1402 is tough enough and is called a low-molecular-weight *polyethylene*, the simplest of all thermoplastics. Continuing the addition of more CH_2 groups further increases the toughness of the material until medium-molecular-weight (between 1000 and 5000 carbons) and

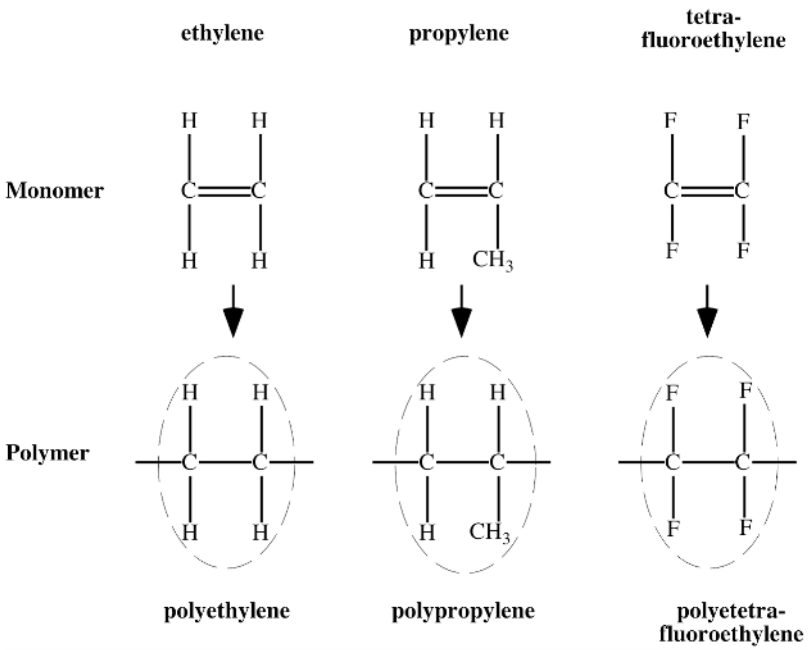


Fig. 18.4. Monomers and their respective polymer units.

high-molecular-weight polyethylene. Polyethylene, being the simplest polymer (Fig. 18.4), has many useful properties in sensor technologies. For example, polyethylene is reasonably transparent in the mid- and far-infrared spectral ranges and thus is used for fabrication of infrared windows and lenses.

By applying heat, pressure, and catalysts, monomers are grown into long chains. The process is called polymerization. Chain length (molecular weight) is important because it determines many properties of a plastic. The major effect of increased length are increased toughness, creep resistance, stress-crack resistance, melt temperature, melt viscosity, and difficulty of processing. After polymerization is completed, the finished polymer chains resemble long intertwined bundles of spaghetti with no physical connections between chains. Such a polymer is called *thermoplastic* (heat-moldable) polymer.

If chains are packed closer to one another, a denser polyethylene is formed which, in effect, results in the formation of crystals. Crystallized areas are stiffer and stronger. Such polymers are more difficult to process because they have higher and sharper melt temperatures; that is, instead of softening, they quickly transform into low-viscosity liquids. On the other hand, amorphous thermoplastics soften gradually, but they do not flow as easily as crystalline plastics. The examples of amorphous polymers are acrylonitrile–butadiene–styrene, polystyrene, polycarbonate, polysulfone, and polyetherimide. Crystalline plastics include polyethylene, polypropylene, nylon, polyvinylidene fluoride (PVDF), acetal, and others.

The following is a nonexhaustive list of thermoplastics:

ABS (acrylonitrile–butadiene–styrene) is very tough, yet hard and rigid. It has fair chemical resistance, low water absorption, and good dimensional stability. Some grades may be electroplated.

Acrylic has high optical clarity and excellent resistance to outdoor weathering. This is a hard, glossy material with good electrical properties. It is available in a variety of colors.

Fluoroplastics comprise a large family of materials (PTFE, FEP, PFA, CTFE, ECTFE, ETFE, and PFDF) with excellent electrical properties and chemical resistance, low friction, and outstanding stability at high temperatures. However, their strength is moderate and the cost is high.

Nylon (polyimide) has outstanding toughness and wear resistance with a low coefficient of friction. It has good electrical and chemical properties. However, it is hygroscopic and dimensional stability is worst than in most other plastics.

Polycarbonate has the highest impact resistance. It is transparent with excellent outdoor stability and resistance to creep under load. It may have some problems with chemicals.

Polyester has excellent dimensional stability but is not suitable for outdoor use or for service in hot water.

Polyethylene is lightweight and inexpensive with excellent chemical stability and good electrical properties. It has moderate transparency in the broad spectral range from visible to far infrared; it has poor dimensional and thermal stability.

Polypropylene has outstanding resistance to flex and stress cracking with excellent chemical and electrical properties with good thermal stability. It is lightweight and inexpensive. Optical transparency is good down to the far-infrared spectral range. However, absorption and scattering of photons in the mid-infrared range is higher than in polyethylene.

Polyurethane is tough, extremely abrasion, and impact resistant. It can be made into films and foams. It has good chemical and electrical properties; however, UV exposure degrades its quality.

Another type of plastic is called *thermoset*, in which polymerization (curing) is done in two stages: one by the material manufacturer and the other by the molder. An example is phenolic, which during the molding process is liquefied under pressure, producing a cross-linking reaction between molecular chains. After it has been molded, a thermoset plastic has virtually all of its molecules interconnected with strong physical bonds, which are not heat reversible. In effect, curing, a thermoset is like cooking an egg. Once it is cooked, it will remain hard. In general, thermoset plastics resist higher temperatures and provide greater dimensional stability. This is the reason why such thermoset plastics such as polyester (reinforced) is used to make boat hulls and circuit-breaker components, epoxy is used to make printed circuit boards, and melamine is used to make dinnerware. On the other hand, thermoplastics offer higher impact strength, easier processing, and better adaptability to complex designs than do thermosets.

The thermoplastics that are most useful in sensor-related applications are the following.

Alkyd has excellent electrical properties and very low moisture absorption.
Allyl (diallyl phtalate) has outstanding dimensional stability and high heat and chemical resistance.
Epoxy has exceptional mechanical strength, electrical properties, and adhesion to most of materials.
Phenolic is a low-cost material. The color is limited to black and brown.
Polyester (thermoplastic version) has a great variety of colors and may be transparent or opaque. Shrinkage is high.

If two different monomers (A and B) are combined in a polymerization reaction, such a polymer is called *copolymer*. The final properties of a copolymer depend on the ratio of components A and B. Polymer mechanical properties can be modified by providing additives, such as fibers to increase strength and stiffness, plastisizers for flexibility, lubricants for easier molding, or UV stabilizers for better performance in sunlight.

Another good way to control properties of plastics is to make polymer alloys or blends. Primarily this is done to retain properties of each component.

Conductive plastics. Being a wonderful electrical isolators, plastic materials often require lamination with metal foil, painting with conductive paint, or metallization to give them electrical conductive properties, required for shielding. Another way of providing electrical conductivity is mixing plastics with conductive additives (e.g., graphite or metal fibers) or building composite plastic parts incorporating metal mesh.

Piezoelectric plastics. These are made from PVF_2 , PVDF, and copolymers which are crystalline materials. Initially, they do not possess piezoelectric properties and must be poled either in high voltage or by corona discharge (Section 3.6 of Chapter 3). Metal electrodes are deposited on both sides of the film either by silkscreening or vacuum metallization. These films, in some applications are used instead of ceramics, because of their flexibility and stability against mechanical stress. Another advantage of the piezoelectric plastics is their ability to be formed into any desirable shape.

18.1.3 Metals

From the sensor designer standpoint, there are two classes of metal: nonferrous and ferrous. Ferrous metals, like steel, are often used in combination with magnetic sensors to measure motion, distance, magnetic field strength, and so forth. Also, they are quite useful as magnetic shields. Nonferrous metals, on the other hand, are permeable to magnetic fields and used whenever these fields are of no concern.

Nonferrous metals offer a wide variety of mechanical and electrical properties. When selecting a metal, one must consider not only its physical properties but also ease of mechanical processing. For example, copper has excellent thermal and electrical properties, yet it is difficult to machine; therefore, in many instances, aluminum

should be considered as a compromise alternative. *Aluminum* has a high strength-to-weigh ratio and possesses its own anticorrosion mechanism. When exposed to air, aluminum does not oxide progressively, like iron would do. The protection is provided by a microscopic oxide coating which forms on the surface and seals the bare metal from the environment.

There are hundreds of aluminum alloys. They can be processed in many ways, such as drawing, casting, and stamping. Some alloys can be soldered and welded. In addition to excellent electrical properties, aluminum is a superb reflector of light over nearly the entire spectrum from UV to radio waves. Aluminum coatings are widely used for mirrors and waveguides. In the mid- and far-infrared range, the only superior to aluminum reflector is gold.

Beryllium has several remarkable properties. Its low density (two-thirds that of aluminum) is combined with a high modulus per weight (five times that of steel), high specific heat, excellent dimensional stability, and transparency to X-rays. However, this is an expensive metal. Like aluminum, beryllium forms a protective coating on its surface, thus resisting corrosion. It may be processed by many conventional methods, including powder cold pressing. The metal is used as X-ray windows, optical platforms, mirror substrates, and satellite structures.

Magnesium is a very light metal with a high strength-to-weight ratio. Due to its low modulus of elasticity, it can absorb energy elastically, which gives its good damping characteristics. The material is very easy to process by most of metal-working techniques.

Nickel allows the design of very tough structures which are also resistant to corrosion. When compared with steel, the nickel alloys have ultrahigh strength and a high modulus of elasticity. Its alloys include binary systems with copper, silicon, and molybdenum. Nickel and its alloys preserve their mechanical properties down to cryogenic temperatures and at high temperatures up to 1200°C. Nickels is used in high-performance superalloys such as Inconell, Monel (Ni–Cu), Ni–Cr, and Ni–Cr–Fe alloys.

Copper combines very good thermal and electrical conductivity properties (second only to pure silver) with corrosion resistance and relative ease of processing. However, its strength-to-weight ratio is relatively poor. Copper is also difficult to machine. Copper and its alloys—the brasses and bronzes—come in variety of forms, including films. Brasses are alloys which contain zinc and other designated elements. Bronzes comprise several main groups: copper–tin–phosphorus (phosphor bronze), copper–tin–lead–phosphorus (lead phosphor bronzes), and copper–silicon (silicon bronzes) alloys. Under outdoor condition, copper develops a blue-green patina. This can be prevented by applying an acrylic coating. A copper alloy with beryllium has excellent mechanical properties and used to make springs.

Lead is the most impervious of all common metals to X-rays and γ- radiation. It resists attack by many corrosive chemicals, most types of soil, and marine and industrial environments. It has a low melting temperature, ease of casting and forming, and good sound and vibration absorption. It possesses natural lubricity and wear resistance. Lead is rarely used in pure form. Its most common alloys are "hard lead" (1–13% of antimony), calcium, and tin alloys which have better strength and hardness.

Platinum is a silver-white precious metal which is extremely malleable, ductile, and corrosion resistant. Its positive temperature coefficient of resistance is very stable and reproducible, which allows its use in temperature sensing.

Gold is extremely soft and chemically inert metal. It can only be attacked by *aqua regia* and by sodium and potassium in the presence of oxygen. One gram of pure gold can be worked into a leaf covering 5000 cm^2 and only less than 0.1 μm thick. Mainly, it is used for plating and is alloyed with other metals like copper, nickel, and silver. In sensor applications, gold is used for fabricating electrical contacts and plating mirrors and waveguides operating in the mid- and far-infrared spectral ranges.

Silver is the least costly of all precious metals. It is very malleable and corrosion resistant. It has the highest electrical and thermal conductivity of all metals.

Palladium, iridium, and *rhodium* resemble and behave like platinum. They are used as electrical coatings to produce hybrid and printed circuit boards and various ceramic substrates with electrical conductors. Another application is in the fabrication of high-quality reflectors operating in a broad spectral range, especially at elevated temperatures or highly corrosive environments. Iridium has the best corrosion resistance of all metals and thus used in the most critical applications.

Molybdenum maintains its strength and rigidity up to 1600°C. The metal and its alloys are readily machinable by conventional tools. In nonoxidizing environments, it resists attacks by most acids. Its prime application is for high-temperature devices, such as heating elements and reflectors of intense infrared radiation for high-temperature furnaces. Molybdenum has a low coefficient of thermal expansion and resists erosion by molten metals.

Tungsten is many respects is similar to molybdenum, but can operate even at higher temperatures. A thermocouple sensor fabricated of tungsten is alloyed with 25% rhenium with another wire, in a thermocouple with 5% rhenium.

Zinc is seldom used alone, except for coating; it is mainly used as an additive in many alloys.

18.1.4 Ceramics

In sensor technologies, ceramics are very useful crystalline materials because of their structural strength, thermal stability, light weight, resistance to many chemicals, ability to bond with other materials, and excellent electrical properties. Although most metals form at least one chemical compound with oxygen, only a handful of oxides are useful as the principal constituent of ceramics. Examples are alumina and beryllia. The natural alloying element in alumina is silica; however, alumina can be alloyed with chromium, magnesium, calcium, and other elements.

Several metal carbides and nitrades qualify as ceramics. The most commonly used are boron carbide and nitrate and aluminum nitrade (Table A.24). Whenever fast heat transfer is of importance, aluminum nitrade should be considered, whereas silicon carbide has high dielectric constant, which makes it attractive for designing capacitive sensors. Due to their hardness, most ceramics require special processing. A precise and cost-effective method of cutting various shapes of ceramic substrates is scribing, machining, and drilling by use of computer-controlled CO_2 laser. Ceramics for the

sensor substrates are available from many manufacturers in thicknesses ranging from 0.1 to 10 mm.

18.1.5 Glasses

Glass is an amorphous solid material made by fusing silica with a basic oxide. Although its atoms never arrange themselves into crystalline structure, the atomic spacing in glass is quite tight. Glass is characterized by transparency, availability in many colors, hardness, and resistance to most chemicals except hydrofluoric acid (Table A.25). Most glasses are based on the silicate system and is made from three major components: silica (SiO), lime ($CaCO_3$), and sodium carbonite ($NaCO_3$). Nonsilicate glasses include phosphate glass (which resists hydrofluoric acid), heat-absorbing glasses (made with FeO), and systems based on oxides of aluminum, vanadium, germanium, and other metals. An example of such specialty glass is arsenic trisulfate (As_2S_3) known as AMTIR, which is substantially transparent in mid- and far-infrared spectral ranges and is used for fabricating infrared optical devices.[2]

Borosilicate glass is the oldest type of glass which is substantially resistant to thermal shock. Under the trademark Pyrex®, some of the SiO_2 molecules are replaced by boric oxide. The glass has a low coefficient of thermal expansion and thus is used for the fabrication optical mirrors (such as in telescopes).

Lead–alkali glass (lead glass) contains lead monoxide (PbO) which increases its index of refraction. Also, it is a better electrical insulator. In the sensor technologies, it is used for fabricating optical windows and prisms and as a shield against nuclear radiation. Other glasses include alumosilicate glass (in which Al_2O_3 replaces some silica), 96% silica, and fused silica.

Another class of glass is *light-sensitive* glasses which are available in three grades. Photochromatic glass darkens when exposed to UV radiation and clears when the UV radiation is removed or glass is heated. Some photochromatic compositions remain darkened for a week or longer. Others fade within few minutes when UV radiation is removed. The photosensitive glass reacts to UV radiation in a different manner: If it is heated after exposure, it changes from clear to opal. This allows the creation of some patterns within the glass structure. Moreover, the exposed opalized glass is much more soluble in hydrofluoric acid, which allows for an efficient etching technique.

18.2 Surface Processing

18.2.1 Deposition of Thin and Thick Films

Thin films are required to give a sensing surface some properties which it otherwise does not possess. For example, to enhance the absorption of thermal radiation by a far-infrared sensor, the surface may be coated with a material having high absorptivity, (e.g., nichrome). A piezoelectric film may be applied to a silicon wafer to give it piezoelectric properties. The thick films are often used to fabricate pressure sensors or microphones where the flexible membranes have to be produced. Several methods

[2] AMTIR infrared glasses are available from Amorphous Materials, Inc. Garland, TX.

may be used to deposit thin and relatively thin (often referred to as "thick") layers of films on a substrate or semiconductor wafer. Among them, the frequently used are the spin-casting, vacuum deposition, sputtering, electroplating, and screenprinting.

18.2.2 Spin-Casting

The spin-casting process involves the use of a thin-film material dissolved in a volatile liquid solvent. The solution is pored on the sample and the sample is rotated at a high speed. The centrifugal forces spread the material, and after the solvent evaporates, a thin layer of film remains on the sample. This techniques is often used for the deposition of organic materials, especially for fabricating humidity and chemical sensors. The thickness depends on the solubility of the deposited material and the spin film and typically is in the range from 0.1 to 50 μm. Because the process relies on the flow of the solution, it may not yield a uniform film or can form island (film-free areas) when the sample has a nonflat surface. In addition, the material may have tendency to shrink. Nevertheless, in many cases, it is a useful and often the only acceptable method of deposition

18.2.3 Vacuum Deposition

A metal can be converted into gaseous form and then deposited on the surface of the sample. The evaporation system consists of a vacuum chamber (Fig. 18.5) where a diffuse pump evacuates air down to 10^{-6}–10^{-7} torr of pressure. A deposited material is placed into a ceramic crucible which is heated by a tungsten filament above the metal melting point. An alternative method of heating is the use of an electron beam.

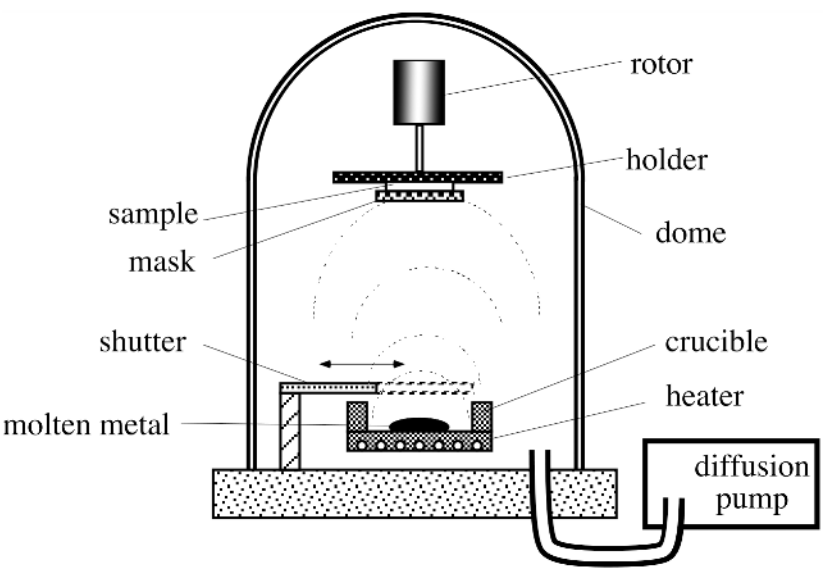


Fig. 18.5. Deposition of a thin metal film in a vacuum chamber.

On a command from the control device, the shutter opens and allows the metal atoms emanated from the molten metal to deposit on the sample. Parts of the sample which remain free of the film are protected by the mask. The film thickness is determined by the evaporation time and the vapor pressure of the metal. Hence, materials with a low melting point are easy to deposit (e.g., aluminum). In general, vacuum-deposited films have large residual stress and thus this technique is used mainly for depositing only thin layers.

Because the molten material is virtually a point source of atoms, it may cause both nonuniform distribution of the deposited film and the so-called shadowing effect where the edges of the masked pattern appear blurry. Two methods may help to alleviate this problem. One is the use of multiple sources where more than one crucible (often three or four) is used. Another method is the rotation of the target.

When using vacuum deposition, one must pay attention to the introduction of spurious materials into the chamber. For instance, even a miniscule amount of oil leaking from the diffuse pump will result in the burning of organic materials and codeposition on the sample of such undesirable compounds as carbohydrates.

18.2.4 Sputtering

As in the vacuum-deposition method, sputtering is performed in a vacuum chamber (Fig. 18.6); however, after evacuation of air, an inert gas, such as argon or helium, is introduced into the chamber at about 2×10^{-6} to 5×10^{-6} torr. An external high-voltage dc or ac power supply is attached to the cathode (target), which is fabricated of the material which has to be deposited on the sample. The sample is attached to the anode at some distance from the cathode. A high voltage ignites the plasma of the inert

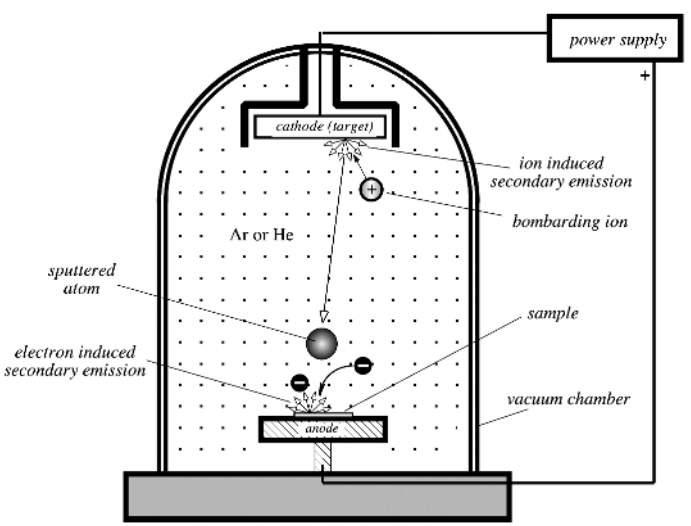


Fig. 18.6. Sputtering process in a vacuum chamber.

gas, and the gas ions bombard the target. The kinetic energy of the bombarding ions is sufficiently high to free some atoms from the target surface. Hence, the escaped sputtered atoms deposit on the surface of the sample.

The sputtered techniques yields better uniformity, especially if a magnetic field is introduced into the chamber, allowing for better directing of the atoms toward the anode. Because this method does not require a high temperature of the target, virtually any material, including organic, can be sputtered. Moreover, materials from more than one target can be deposited at the same time (cosputtering), permitting a controlled ratio of materials. For example, this can be useful for sputtering nichrome (Ni and Cr) electrodes on the surface of the pyroelectric sensors.

18.2.5 Chemical Vapor Deposition

A chemical vapor phase deposition (CVD) process is an important technique for the production of optical, optoelectronic, and electronic devices. For sensor technologies, it is useful for forming optical windows and the fabrication of semiconductor sensors where thin and thick crystalline layers have to be deposited on the surface.

The CVD process takes place in a deposition (reaction) chamber, one of the versions of which is shown in a simplified form in Fig. 18.7. The substrates or wafers are positioned on a stationary or rotating table (the substrate holder) whose temperature is elevated up to the required level by the heating elements. The top cover of the chamber has an inlet for the carrier H_2 gas, which can be added by various precursors and dopants. These additives, while being carried over the heated surface of the substrate, form a film layer. The gas mixture flows from the distribution cone over the top surface of the wafers and exits through the exhaust gas outlets. The average gas pressure in the chamber may by near 1 atm, or somewhat lower. For example, a

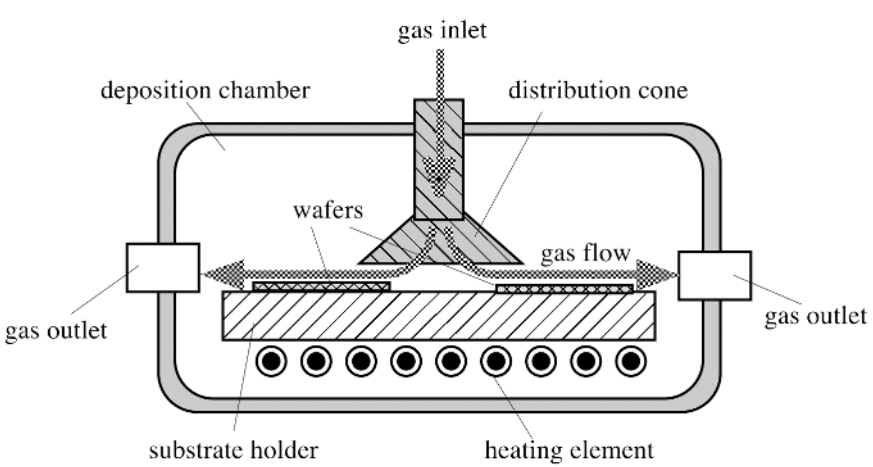


Fig. 18.7. Simplified structure of a CVD reactor chamber.

6000-Å layer of $Ga_{0.47}In_{0.53}As$ can be grown on the InP substrate at 1 atm and 630°C with a rate of 1.4 Å/s [3].

18.3 Nano-Technology

Nano-technology today is a somewhat emotional term, more of a wishful thinking than a real thing. It refers to dimensions of a device comparable with a nanometer (10^{-9} m) scale. In practice, however, most of the subminiature elements have sizes about 1000 times larger—in a micrometer (10^{-6} m) range. Still, the trend is toward the smallest dimensions as far as the current technology allows.

The present trend in sensor technologies is undoubtedly shifted toward the micro miniaturization or *microsystem technologies*, known as MST. A subset of these is known as *micro-electromechanical systems*, or MEMS for short. A MEMS device has electrical and mechanical components, which means there must be at least one moving or deformable part and that electricity must be part of its operation. Another subset is called MEOMS, which stands for micro-electro-optical systems. As the name implies, at least one optical component is part of the device. Most of the sensors that are fabricated with the use of MEMS or MEOMS are three-dimensional devices with dimensions on the order of micrometers.

The two constructional technologies of microengineering are *microelectronics* and *micromachining*. Microelectronics, producing electronic circuitry on silicon chips, is a very well-developed technology. Micromachining is the name for the techniques used to produce the structures and moving parts of microengineered devices. One of the main goals of microengineering is to be able to integrate microelectronic circuitry into micromachined structures, to produce completely integrated systems (microsystems). Such systems typically have the same advantages of low cost, reliability, and small size as silicon chips produced in the microelectronics industry.

Presently, there are three micromachining techniques that are in use or are extensively developed by the industry [4,5]. *Silicon micromachining* is given the most prominence, because this is one of the better developed micromachining techniques. Silicon is the primary substrate material used in the production microelectronic circuitry and, thus, is the most suitable candidate for the eventual production of microsystems.

The *excimer laser* is an ultraviolet laser which can be used to micromachine a number of materials without heating them, unlike many other lasers which remove material by burning or vaporizing it. The excimer laser lends itself particularly to the machining of organic materials (polymers, etc).

LIGA[3] is a technique that can be used to produce molds for the fabrication of micromachined components. Microengineered components can be made from a variety of materials using this technique, however it does suffer the disadvantage that the technique currently requires X-rays from a synchrotron source.

[3] LIGA-Lithographic Galvanoforming and Abforming is a German acronym for x-ray lithography.

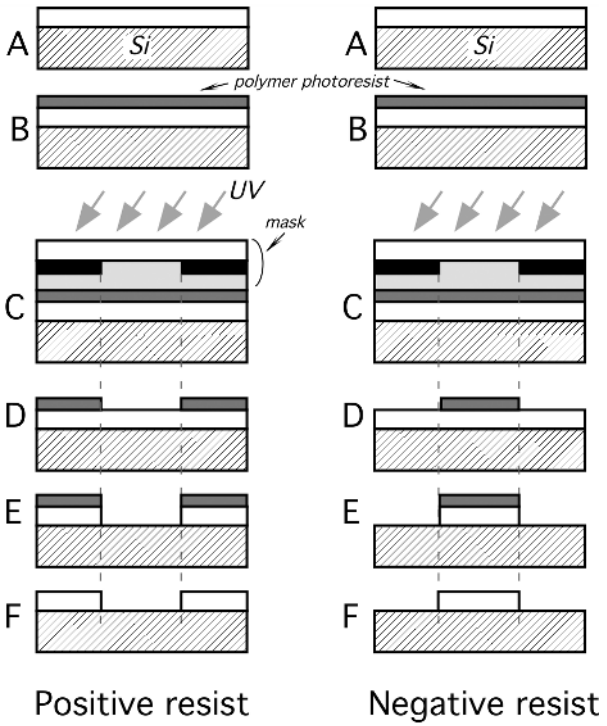


Fig. 18.8. Positive and negative photolithography.

18.3.1 Photolithography

Photolithography is the basic technique used to define the shape of micromachined structures in the three techniques outlined below. The technique is essentially the same as that used in the microelectronics industry.

Figure 18.8A shows a thin film of some material (e.g., silicon dioxide) on a substrate of some other material (e.g., a silicon wafer). The goal of the process is to selectively remove some silicon dioxide (oxide) so that it only remains in particular areas on the silicon wafer (Fig. 18.8F).

The process beings with producing a mask. This will typically be a chromium pattern on a glass plate. The wafer is then coated with a polymer which is sensitive to UV light (Fig. 18.8B), called a photoresist. Ultraviolet light is then shone through the mask onto the photoresist (Fig. 18.8C). The photoresist is then developed which transfers the pattern on the mask to the photoresist layer (Fig. 18.8D).

There are two types of photoresist, termed positive (left side of Fig. 18.8) and negative (right side of Fig. 18.8). Where the ultraviolet light strikes the positive resist, it weakens the polymer, so that when the image is developed, the resist is washed away where the light struck it—transferring a positive image of the mask to the resist layer. The opposite occurs with negative resist. Where the ultraviolet light strikes negative resist it strengthens the polymer, so when developed, the resist that was not

exposed to UV light is washed away—a negative image of the mask is transferred to the resist. A chemical (or some other method) is then used to remove the oxide where it is exposed through the openings in the resist (Fig. 18.8E). Finally, the resist is removed, leaving the patterned oxide (Fig. 18.8F).

18.3.2 Silicon Micromachining

There is a number of basic techniques that can be used to pattern thin films that have been deposited on a silicon wafer and to shape the wafer itself to form a set of basic microstructures (bulk silicon micromachining). The techniques for depositing and patterning thin films can be used to produce quite complex microstructures on the surface of silicon wafer (surface silicon micromachining). Electrochemical *etching* techniques are being investigated to extend the set of basic silicon micromachining techniques. Silicon bonding techniques can also be utilized to extend the structures produced by silicon micromachining techniques into multilayer structures.

18.3.2.1 Basic Techniques

There are three basic techniques associated with silicon micromachining. These are the deposition of thin films of materials, the removal of material (patterning) by wet chemical etchants, and the removal of material by dry-etching techniques. Another technique that is utilized is the introduction of impurities into the silicon to change its properties (i.e., doping).

18.3.2.1.1 Thin Films

There are a number of different techniques that facilitate the deposition or formation of very thin films (of the order of micrometers or less) of different materials on a silicon wafer (or other suitable substrate). These films can then be patterned using photolithographic techniques and suitable etching techniques. Common materials include silicon dioxide (oxide), silicon nitride (nitride), polycrystalline silicon (polysilicon or poly), and aluminum. A number of other materials can be deposited as thin films, including noble metals such as gold. However, noble metals will contaminate microelectronic circuitry causing it to fail, so any silicon wafers with noble metals on them have to be processed using equipment specially set aside for the purpose. Noble metal films are often patterned by a method known as "lift off," rather than wet or dry etching.

Often, photoresist is not tough enough to withstand the etching required. In such cases, a thin film of a tougher material (e.g., oxide or nitride) is deposited and patterned using photolithography. The oxide/nitride then acts as an etch mask during the etching of the underlying material. When the underlying material has been fully etched, the masking layer is stripped away.

18.3.2.1.2 Wet Etching

Wet etching is a blanket name that covers the removal of material by immersing the wafer in a liquid bath of the chemical etchant. Wet etchants fall into two broad categories: isotropic etchants and anisotropic etchants. Isotropic etchants attack the

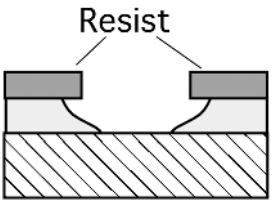


Fig. 18.9. Isotropic etching under the mask.

material being etched at the same rate in all directions. Anisotropic etchants attack the silicon wafer at different rates in different directions, and so there is more control of the shapes produced. Some etchants attack silicon at different rates depending on the concentration of the impurities in the silicon (concentration-dependent etching).

Isotropic etchants are available for oxide, nitride, aluminum, polysilicon, gold, and silicon. Because isotropic etchants attack the material at the same rate in all directions, they remove material horizontally under the etch mask (undercutting) at the same rate as they etch through the material. This is illustrated for a thin film of oxide on a silicon wafer in Fig. 18.9, using an etchant that etches the oxide faster than the underlying silicon (e.g., hydroflouric acid).

Anisotropic etchants are available to etch different crystal planes in silicon at different rates. The most popular anisotropic etchant is potassium hydroxide (KOH), because it is the safest to use.

Etching is done on a *silicon wafer*. Silicon wafers are slices that have been cut from a large ingot of silicon that was grown from a single seed crystal. The silicon atoms are all arranged in a crystalline structure, so the wafer is monocrystalline silicon (as opposed to polycrystalline silicon mentioned earlier). When purchasing silicon wafers, it is possible to specify that they have been sliced with the surface parallel to a particular crystal plane.

The simplest structures that can be formed using KOH to etch a silicon wafer with the most common crystal orientation (100) are shown in Fig. 18.10. These are the V-shaped grooves or pits with right-angled corners and sloping side walls. Using wafers with different crystal orientations can produce grooves or pits with vertical walls.

Both oxide and nitride etch slowly in KOH. Oxide can be used as an etch mask for short periods in the KOH etch bath (i.e., for shallow grooves and pits). For long periods, nitride is a better etch mask, as it etches more slowly in the KOH.

KOH can also be used to produce mesa structures (Fig. 18.11A). When etching mesa structures, the corners can become beveled (Fig. 18.11B), rather than right-angle corners. This has to be compensated for in some way. Typically, the etch mask is designed to include additional structures on the corners. These compensation structures are designed so that they are etched away entirely when the mesa is formed to leave 90° corners. One problem with using compensation structures to from right-angle mesa corners is that they put a limit on the minimum spacing between the mesas.

Fabrication of a diaphragm is one of the most popular sensor processes. It is used to produce accelerometers, pressures sensor, infrared temperature sensors (thermopiles and bolometers), and many others. Silicon diaphragms from about 50 μm thick upward can be made by etching through an entire wafer with KOH (Fig. 18.12A). The thickness is controlled by timing the etch and thus is subject to errors.

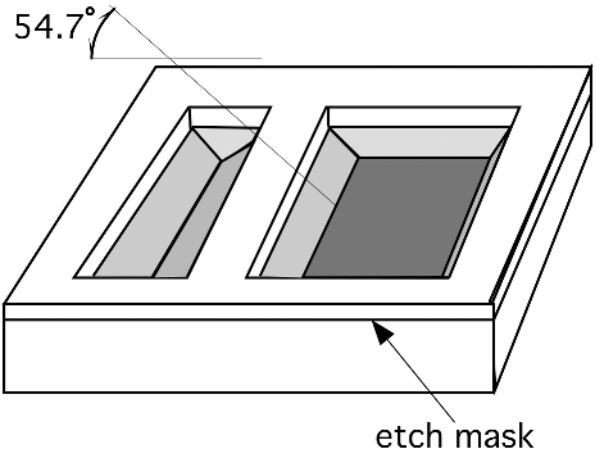


Fig. 18.10. Simple structures etched by KOH.

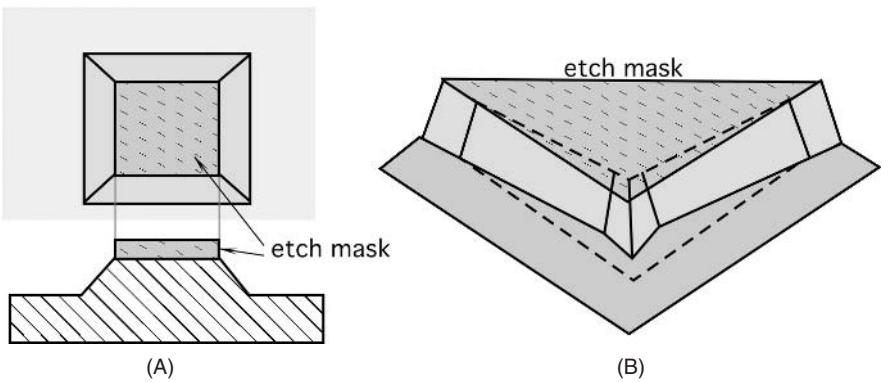


Fig. 18.11. Mesa structures.

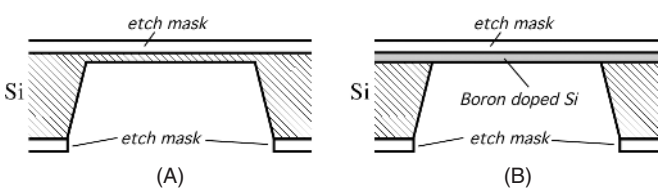


Fig. 18.12. Micromachining of a diaphragm or membrane.

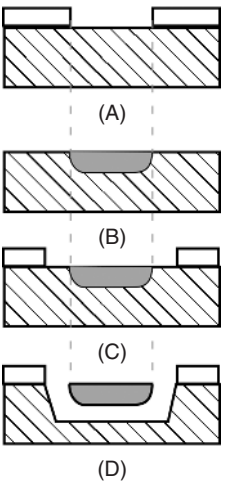


Fig. 18.13. Etching around the boron-doped silicon.

18.3.2.1.3 Concentration-Dependent Etching

Thinner diaphragms, up to about 20 μm thick, can be produced using boron to stop the KOH etch (Fig. 18.12B). This is called the concentration-dependent etching. The thickness of the diaphragm is dependent on the depth to which the boron is diffused into the silicon, which can be controlled more accurately than the simple, timed KOH etch. High levels of boron in silicon will reduce the rate at which it is etched in KOH by several orders of magnitude, effectively stopping the etching of the boron-rich silicon. The boron impurities are usually introduced into the silicon by a process known as *diffusion*.

In addition to the diaphragms, many other structures can be built by the concentration-dependent etching. A thick oxide mask is formed over the silicon wafer and patterned to expose the surface of the silicon wafer where the boron is to be introduced (Fig. 18.13A). The wafer is then placed in a furnace in contact with a boron diffusion source. Over a period of time, boron atoms migrate into the silicon wafer. Once the boron diffusion is completed, the oxide mask is stripped off (Fig. 18.13B). A second mask may then be deposited and patterned (Fig. 18.13C) before the wafer is immersed in the KOH etch bath. The KOH etches the silicon that is not protected by the mask, and it etches around the boron-doped silicon (Fig. 18.13D). Boron can be driven into the silicon as far as 20 μm over periods of 15–20 h; however, it is desirable to keep the time in the furnace as short as possible. Concentration-dependent etching can also be used to produce narrow bridges or cantilever beams. Figure 18.14A shows a bridge, defined by a boron diffusion, spanning a pit that was etched from the front of the wafer in KOH. A cantilever beam (a bridge with one end free) produced by the same method is shown in Fig. 18.14B. The bridge and beam project across the diagonal of the pit to ensure that they will be etched free by the KOH. More complex structures are possible using this technique, but care must be taken to ensure that they will be etched free by the KOH.

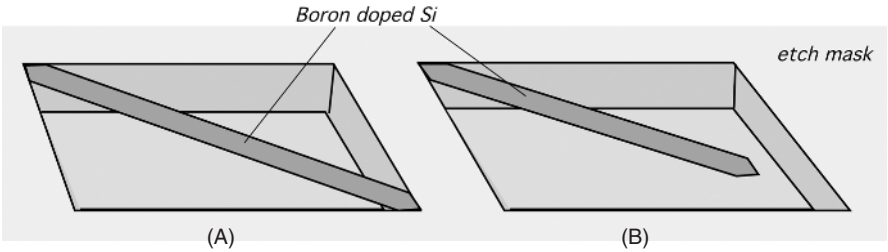


Fig. 18.14. Etching of a bridge and cantilever.

One of the applications for these beams and bridges is the resonant sensors. The structure can be set vibrating at its fundamental frequency. Anything causing a change in the mass, length, and so forth, of the structure will register as a change frequency. Care has to be taken to ensure that only the quantity to be measured causes a significant change in frequency.

18.3.2.1.4 Dry Etching

The most common form of dry etching for micromachining applications is *reactive ion etching* (RIE). Ions are accelerated toward the material to be etched, and the etching reaction is enhanced in the direction of travel of the ion. RIE is an anisotropic etching technique. Deep trenches and pits (up to ten or a few tens of microns) of arbitrary shape and with vertical walls can be etched in a variety of materials, including silicon, oxide, and nitride. Unlike anisotropic wet etching, RIE is not limited by the crystal planes in the silicon. A combination of dry etching and isotropic wet etching can be used to form very sharp points. First, a column with vertical sides is etched away using an RIE (Fig. 18.15A). A wet etch is then used, which undercuts the etch mask, leaving a very fine point (Fig. 18.15B); the etch mask is then removed. Very fine points like this can be fabricated on the end of cantilever beams as probes for use, for example, in tactile sensors.

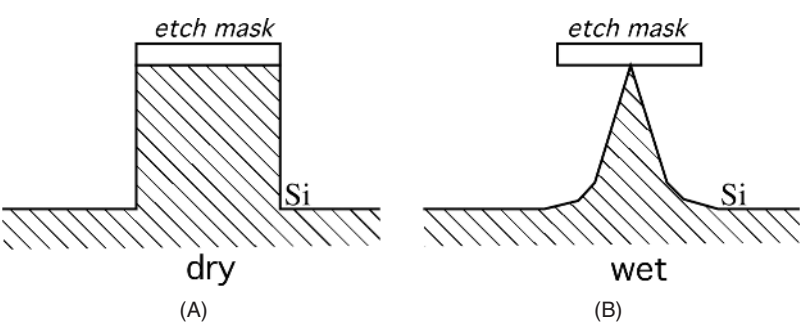


Fig. 18.15. Dry etching of a pointed structure.

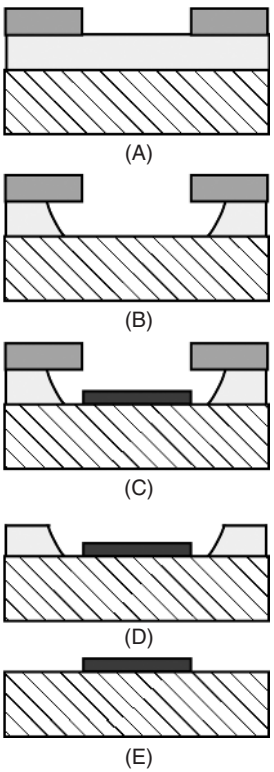


Fig. 18.16. Lift-off technique.

18.3.2.1.5 Lift-Off

Lift-off is a stenciling technique often used to pattern noble metal films. There are a number of different techniques; the one outlined here is an assisted lift off method. A thin film of the assisting material (e.g., oxide) is deposited. A layer of resist is put over this and patterned, as for photolithography, to expose the oxide in the pattern desired for the metal (Fig. 18.16A). The oxide is then wet etched so as to undercut the resist (Fig. 18.16B). The metal is then deposited on the wafer, typically by a process known as evaporation (Fig. 18.16C). The metal pattern is effectively stenciled through the gaps in the resist, which is then removed, lifting off the unwanted metal with it (Fig. 18.16D). The assisting layer is then stripped off too, leaving the metal pattern alone (Fig. 18.16E).

18.3.2.2 Wafer bonding

There are a number of different methods available for bonding micromachined silicon wafers together, or to other substrates, to form larger more complex devices. A method of bonding silicon to glass that appears to be gaining in popularity is anodic bonding (electrostatic bonding). The silicon wafer and glass substrate are brought together and

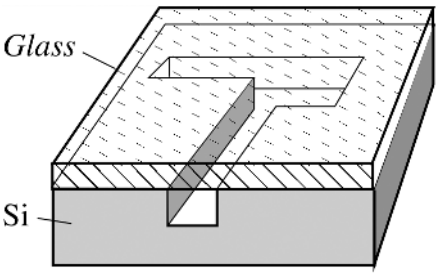


Fig. 18.17. Bonding of glass to silicon.

heated to a high temperature. A large electric field is applied across the join, which causes an extremely strong bond to form between the two materials. Figure 18.17 shows a glass plate bonded over a channel etched into a silicon wafer (RIE).

It is also possible to bond silicon wafers directly together using gentle pressure, under water (direct silicon bonding). Other bonding methods include using an adhesive layer, such as a glass, or photoresist. Although anodic bonding and direct silicon bonding form very strong joins, they suffer from some disadvantages, including the requirement that the surfaces to be joined are very flat and clean. Wafer bonding techniques can potentially be combined with some of the basic micromachined structures to form the membranes, cantilevers, valves, pumps, and so forth, of a microfluid handling system that may be parts of chemical sensors.

References

1. Middelhoek, S. and Hoogerwerf A.C. Smart sensors: when and where? *Sensors Actuators* 8(1), 39–48, 1985.
2. Obermier, E., Kopystynski, P. and Neißl, R. Characteristics of polysilicon layers and their application in sensors. IEEE Solid-State Sensors Workshop, 1986.
3. Frijlink, P.M., Nicolas, J.L., and Suchet, P. Layer uniformity in a multiwafer MOVPRE reactor for III–V compounds. *J. Crystal Growth* 107, 167–174, 1991.
4. Morgan, D.V. and Board, K. *An Introduction to Semiconductor Microtechnology*, John Wiley & Sons, New York, 1985.
5. Muller, R.S., Howe, R.T., Senturia, S.D., Smith, R.L., and White R.M. (eds.). *Microsensors*, IEEE Press, New York, 1991.

This page intentionally left blank

Appendix

Table A.1. Chemical Symbols for the Elements

Ac	Actinium	Es	Einsteinium	N	Nitrogen	Sc	Scandium
Ag	Silver	Eu	Europium	Na	Sodium	Se	Selenium
Al	Aluminum	F	Fluorine	Nb	Niobium	Si	Silicon
Am	Americium	Fe	Iron	Nd	Neodymium	Sm	Samarium
Ar	Argon	Fm	Fermium	Ne	Neon	Sn	Tin
As	Arsenic	Fr	Francium	Ni	Nickel	Sr	Strontium
At	Astatine	Ga	Gallium	No	Nobelium	Ta	Tantalum
Au	Gold	Gd	Gadolinium	Np	Neptunium	Tb	Terbium
B	Boron	Ge	Germanium	O	Oxygen	Tc	Technetium
Ba	Barium	H	Hydrogen	Os	Osmium	Te	Tellurium
Be	Beryllium	He	Helium	P	Phosphorous	Th	Thorium
Bi	Bismuth	Hf	Hafnium	Pa	Protactinium	Ti	Titanium
Bk	Berkelium	Hg	Mercury	Pb	Lead	Tl	Thallium
Br	Bromine	Ho	Holmium	Pd	Palladium	Tm	Thulium
C	Carbon	I	Iodine	Pm	Promethium	U	Uranium
Ca	Calcium	In	Indium	Po	Polonium	V	Vanadium
Cd	Cadmium	Ir	Iridium	Pr	Praseodymium	W	Tungsten
Ce	Cerium	K	Potassium	Pt	Platinum	Xe	Xenon
Cf	Californium	Kr	Krypton	Pu	Plutonium	Y	Yttrium
Cl	Chlorine	La	Lanthanum	Ra	Radium	Yb	Ytterbium
Cm	Curium	Li	Lithium	Rb	Rubidium	Zn	Zinc
Co	Cobalt	Lr	Lawrencium	Re	Rhenium	Zr	Zirconium
Cr	Chromium	Lu	Lutetium	Rh	Rhodium		
Cs	Cesium	Md	Mendelevium	Rn	Radon		
Cu	Copper	Mg	Magnesium	Ru	Ruthenium		
Dy	Dysprosium	Mn	Manganese	S	Sulfur		
Er	Erbium	Mo	Molybdenum	Sb	Antimony		

Table A.2. SI Multiples

Factor	Prefix	Symbol	Factor	Prefix	Symbol
10^{18}	exa	E	10^{-1}	deci	d
10^{15}	peta	P	10^{-2}	centi	c
10^{12}	tera	T	10^{-3}	milli	m
10^{9}	giga	G	10^{-6}	micro	μ
10^{6}	mega	M	10^{-9}	nano	n
10^{3}	kilo	k	10^{-12}	pico	p
10^{2}	hecto	h	10^{-15}	femto	f
10^{1}	deka	da	10^{-18}	atto	a

Table A.3. Derivative SI Units

Quantity	Name of Unit	Expression in Terms of Basic Units
Area	Square meter	m^2
Volume	Cubic meter	m^3
Frequency	Hertz (Hz)	s^{-1}
Density (concentration)	Kilogram per cubic meter	kg/m^3
Velocity	Meter per second	m/s
Angular velocity	Radian per second	rad/s
Acceleration	Meter per second squared	m/s^2
Angular acceleration	Radian per second squared	rad/s^2
Volumetric flow rate	Cubic meter per second	m^3/s
Force	Newton (N)	$kg\ m/s^2$
Pressure	Newton per square meter (N/m^2) or pascal (Pa)	$kg/m\ s^2$
Work energy heat torque	Joule (J), newton-meter (N m) or watt-second (W s)	$kg\ m^2/s^2$
Power heat flux	Watt (W), Joule per second (J/s)	$kg\ m^2/s^3$
Heat flux density	Watt per square meter (W/m^2)	kg/s^3
Specific heat	Joule per kilogram degree (J/kg deg)	m^2/s^2 deg
Thermal conductivity	Watt per meter degree (W/m deg) or (J m/s m^2 deg)	$kg\ m/s^3$ deg
Mass flow rate (mass flux)	Kilogram per second	kg/s
Mass flux density	Kilogram per square meter-second	kg/m^2 s
Electric charge	Coulomb (C)	A s
Electromotive force	Volt (V) or W/A	$kg\ m^2/A\ s^3$
Electric resistance	Ohm (Ω) or V/A	$kg\ m^2/A^2\ s^3$
Electric conductivity	Ampere per volt-meter (A/V m)	$A^2\ s^3/kg\ m^3$
Electric capacitance	Farad (F) or A s/V	$A^3\ s^4/kg\ m^2$
Magnetic flux	Weber (Wb) or V s	$kg\ m^2/A\ s^2$
Inductance	Henry (H) or V s/A	$kg\ m^2/A^2\ s^2$
Magnetic permeability	Henry per meter (H/m)	$kg\ m/A^2\ s^2$
Magnetic flux density	Tesla (T) or weber per square meter (Wb/m^2)	$kg/A\ s^2$
Magnetic field strength	Ampere per meter	A/m
Magnetomotive force	Ampere	A
Luminous flux	lumen (lm)	cd sr
Luminance	Candela per square meter	cd/m^2
Illumination	lux (lx) or lumen per square meter (lm/m^2)	$cd\ sr/m^2$

Table A.4. SI Conversion Multiples

		Acceleration (m/s^2)	
ft/s^2	0.3048	gal	0.01
Free fall (g)	9.80665	in/s^2	0.0254
		Angle [radian (rad)]	
Degree	0.01745329	Second	4.848137×10^{-6}
Minute	2.908882×10^{-4}	Grade	1.570796×10^{-2}
		Area: (m^2)	
Acre	4046.873	Hectare	1×10^4
Are	100.00	mi^2 (U.S. statute)	2.589998×10^6
ft^2	9.290304×10^{-2}	yd^2	0.8361274
		Bending Moment or Torque: (N m)	
Dyne cm	1×10^{-7}	lbf in	0.1129848
kgf m	9.806650	lbf ft	1.355818
ozf in	7.061552×10^{-3}		
		Electricity and Magnetism[a]	
Ampere hour	3600 coulomb (C)	EMU of inductance	8.987×10^{11} henry (H)
EMU of capacitance	10^9 farad (F)	EMU of resistance	$8.987 \times 10^{11}\,(\Omega)$
EMU of current	10 ampere (A)	Faraday	9.65×10^{19} coulomb (C)
EMU of elec. potential	10^{-8} volt (V)	Gamma	10^{-9} tesla (T)
EMU of inductance	10^{-9} henry (H)	Gauss	10^{-4} tesla (T)
EMU of resistance	10^{-9} ohm (Ω)	Gilbert	0.7957 ampere (A)
ESU of capacitance	1.112×10^{-12} farad (F)	Maxwell	10^{-8} weber (Wb)
ESU of current	3.336×10^{-10} ampere (A)	mho	1.0 siemens (S)
EMU of elec. potential	299.79 volt (V)	ohm centimeter	0.01 ohm meter (Ω m)

Table A.4 *Continued*

		Energy (Work): [joule (J)]	
British thermal unit (Btu)	1055	Kilocalorie	4187
Calorie	4.18	kW h	3.6×10^6
Calorie (kilogram)	4184	Ton (nuclear equiv. TNT)	4.184×10^9
Electronvolt	1.60219×10^{-19}	therm	1.055×10^8
erg	10^{-7}	W h	3600
ft lbf	1.355818	W s	1.0
ft poundal	0.04214		
		Force [newton (N)]	
Dyne	10^{-5}	Ounce-force	0.278
Kilogram-force	9.806	Pound-force (lbf)	4.448
Kilopond (kp)	9.806	Poundal	0.1382
kip (1000 lbf)	4448	Ton-force (2000 lbf)	8896
		Heat	
Btu ft/(h ft^2 °F) (thermal conductivity)	1.7307 W/(m K)	cal/cm^2	4.18×10^4 J/m^2
Btu/lb	2324 J/kg	cal/(cm^2 min)	697.3 W/m^2
Btu/(lb °F) = cal/(g °C) (heat capacity)	4186 J/(kg K)	cal/s	4.184 W
Btu/ft^3	3.725×10^4 J/m^3	°F h ft^2/Btu (thermal resistance)	0.176 K m^2/W
cal/(cm s °C)	418.4 W/(m K)	ft^2/h (thermal diffusivity)	2.58×10^{-5} m^2/s
		Length [meter (m)]	
Angstrom	10^{-10}	Microinch	2.54×10^{-8}
Astronomical unit	1.495979×10^{11}	Micrometer (micron)	10^{-6}
Chain	20.11	Mil	2.54×10^{-5}
Fermi (femtometre)	10^{-15}	Mile (nautical)	1852.000
Foot	0.3048	Mile (international)	1609.344
Inch	0.0254	Pica (printer's)	4.217×10^{-3}
Light year	9.46055×10^{15}	Yard	0.9144

Table A.4 *Continued*

		Light	
cd/in.^{2}	1550 cd/m^2	lambert	3.183×10^3 cd/m^2
Foot candle	10.76 lx (lux)	lm/ft^2	10.76 lm/m^2
Foot lambert	3.426 cd/m^2		
		Mass [kilogram (kg)]	
Carat (metric)	2×10^{-4}	Ounce (troy or apothecary)	3.110348×10^{-2}
Grain	6.479891×10^{-5}	Pennyweight	1.555×10^{-3}
Gram	0.001	Pound (lb avoirdupois)	0.4535924
Hundred weight (long)	50.802	Pound (troy or apothecary)	0.3732
Hundred weight (short)	45.359	Slug	14.5939
kgf s^2/m	9.806650	Ton (long, 2120 lbs)	907.184
Ounce (avoirdupois)	2.834952×10^{-2}	Ton (metric)	1000
		Mass per Unit Time (Includes Flow)	
perm (0°C)	5.721×10^{-11} kg/(Pa s m^2)	lbs/(hp h) SPC (specific fuel consumption)	1.689659×10^{-7} kg/J
lbs/h	1.2599×10^{-4} kg/s	Ton (short)/h	0.25199 kg/s
lbs/s	0.4535912 kg/s		
		Mass per Unit Volume (Includes Density and Capacity) (kg/m3)	
oz (avoirdupois)/gal (U.K. liquid)	6.236	oz (avoirdupois)/gal (U.S. liquid)	7.489
oz (avoirdupois)/in.^{3}	1729.99	Slug/ft^3	515.3788
lbs/gal (U.S. liquid)	11.9826 kg/m^3	Ton (long)/yd^3	1328.939
		Power: watt (W)	
Btu (International)/s	1055.056	Horsepower (electric)	746
cal/s	4.184	Horsepower (metric)	735.499
erg/s	10^{-7}	Horsepower (U.K.)	745.7
Horsepower (550 ft lbf/s)	745.6999	Ton of refrigeration (12,000 Btu/h)	3517

Table A.4 *Continued*

	Pressure or Stress [pascal (Pa)]		
Atmosphere, standard	1.01325×10^5	Dyne/cm^2	0.1
Atmosphere, technical	9.80665×10^4	Foot of water (39.2°F)	2988.98
Bar	10^5	Poundal/ft^2	1.488164
Centimeter of mercury (0°C)	1333.22	psi (lbf/in.^{2})	6894.757
Centimeter of water (4°C)	98.0638	Torr (mm Hg, 0°C)	133.322
	Radiation Units		
Curie	3.7×10^{10} becquerel (Bq)	Rem	0.01 sievert (Sv)
rad	0.01 gray (Gy)	Roentgen	2.58×10^{-4} C/kg
	Temperature		
°Celsius	$T\,(\mathrm{K}) = t\,(^\circ\mathrm{C}) + 273.15$ K	° Fahrenheit	$T^\circ(\mathrm{C}) = [t\,(^\circ\mathrm{F}) - 32]/1.8^\circ\mathrm{C}$
°Fahrenheit	$T\,(\mathrm{K}) = [t^\circ\mathrm{F}{+}459.67]/1.8$ K	° Rankine	$T\,(\mathrm{K}) = T\,(^\circ\mathrm{R})/1.8$
	Velocity (Includes Speed) (m/s)		
ft/s	0.3048	mi/h (international)	0.44704
in./s	2.54×10^{-2}	rpm (revolutions/min)	0.1047 rad/s
Knot (international)	0.51444		
	Viscosity: (Pa s)		
Centipose (dynamic viscosity)	10^{-3}	lbf s/in.^{2}	6894.757
Centistokes (kinematic viscosity)	10^{-6}	rhe	10 1/(Pa s)
poise	0.1	Slug/(ft s)	47.88026
Poundal s/ft^2	1.488164	Stokes	10^{-4} m^2/s
lbs/(ft s)	1.488164		
	Volume (Includes Capacity) (m^3)		
Acre-foot	1233.489	Gill (U.S.)	1.182941×10^{-4}
Barrel (oil, 42 gal)	0.1589873	in.^{3}	1.638706×10^{-5}
Bushel (U.S.)	3.5239×10^{-2}	Liter	10^{-3}

Table A.4 *Continued*

Cup	2.36588×10^{-4}	Ounce (U.S. fluid)	2.957353×10^{-5}
Ounce (U.S. fluid)	2.95735×10^{-5}	Pint (U.S. dry)	5.506105×10^{-4}
ft^3	2.83168×10^{-2}	Pint (U.S. liquid)	4.731765×10^{-4}
Gallon (Canadian, U.K. liquid)	4.54609×10^{-3}	Tablespoon	1.478×10^{-5}
Gallon (U.S. liquid)	3.7854×10^{-3}	Ton (register)	2.831658
Gallon (U.S. dry)	4.40488×10^{-3}	yd^3	0.76455

[a]ESU means electrostatic cgs unit; EMU means electromagnetic cgs unit.

Table A.5. Dielectric Constants of Some Materials at Room Temperature (25°C)

Material	κ	Frequency (Hz)	Material	κ	Frequency (Hz)
Air	1.00054	0	Paraffin	2.0–2.5	10^6
Alumina ceramic	8–10	10^4	Plexiglas	3.12	10^3
Acrylics	2.5–2.9	10^4	Polyether sulfone	3.5	10^4
ABS/polysulfone	3.1	10^4	Polyesters	3.22–4.3	10^3
Asphalt	2.68	10^6	Polyethylene	2.26	10^3–10^8
Beeswax	2.9	10^6	Polypropylenes	2–3.2	10^4
Benzene	2.28	0	Polyvinyl chloride	4.55	10^3
Carbon tetrachloride	2.23	0	Porcelain	6.5	0
Cellulose nitrate	8.4	10^3	Pyrex glass (7070)	4.0	10^6
Ceramic (titanium dioxide)	14–110	10^6	Pyrex glass (7760)	4.5	0
Cordierite	4–6.23	10^4	Rubber (neoprene)	6.6	10^3
Compound for thick-film capacitors	300–5000	0	Rubber (silicone)	3.2	10^3
Diamond	5.5	10^8	Rutile $\perp$ optic axis	86	10^8
Epoxy resins	2.8–5.2	10^4	Rutile $\parallel$ optic axis	170	10^8
Ferrous oxide	14.2	10^8	Silicone resins	3.4–4.3	10^4
Flesh (skin, blood, muscles)	97	40×10^6	Tallium chloride	46.9	10^8
Flesh (fat, bones)	15	40×10^6	Teflon	2.04	10^3–10^8
Lead nitrate	37.7	6×10^7	Transformer oil	4.5	0
Methanol	32.63	0	Vacuum	1	—
Nylon	3.5–5.4	10^3	Water	78.5	0
Paper	3.5	0			

Table A.6. Properties of Magnetic Materials

Material	MEP [G Oe $\times 10^6$]	Residual Induction [G $\times 10^3$]	Coercive Force (Oe $\times 10^3$)	Temperature Coefficient (%/°C)	Cost
R.E. Cobalt	16	8.1	7.9	−0.05	Highest
Alnico 1, 2, 3, 4	1.3–1.7	5.5–7.5	0.42–0.72	−0.02 to −0.03	Medium
Alnico 5, 6, 7	4.0–7.5	10.5–13.5	0.64–0.78	−0.02 to −0.03	Medium/high
Alnico 8	5.0–6.0	7–9.2	1.5–1.9	−0.01 to 0.01	Medium/high
Alnico 9	10	10.5	1.6	−0.02	High
Ceramic 1	1.0	2.2	1.8	−0.2	Low
Ceramic 2, 3, 4, 6	1.8–2.6	2.9–3.3	2.3–2.8	−0.2	Low/medium
Ceramic 5, 7, 8	2.8–3.5	3.5–3.8	2.5–3.3	−0.2	Medium
Cunife	1.4	5.5	0.53	—	Medium
Fe–Cr	5.25	13.5	0.6	—	Medium/high
Plastic	0.2–1.2	1.4	0.45–1.4	−0.2	Lowest
Rubber	0.35–1.1	1.3–2.3	1–1.8	−0.2	Lowest

Source: Adapted from Sprague, CN-207 Hall Effect IC applications, 1986.

Table A.7. Some Materials at Room Temperature

Material	ρ ($\times 10^{-8}\,\Omega$ m)	TCR (α) ($\times 10^{-3}$/1)	Material	ρ ($\times 10^{-8}\,\Omega$ m)	TCR (α) ($\times 10^{-3}$/1)
Alumina[a]	$> 10^{20}$		Palladium	10.54	3.7
Aluminum (99.99%)	2.65	3.9	Platinum	10.42	3.7
Beryllium	4.0	0.025	Platinum + 10% rhodium	18.2	
Bismuth	10^6		Polycrystalline glass[a]	6.3×10^{14}	
Brass (70Cu, 30Zn)	7.2	2.0	Rare earth metals	28–300	
Carbon	3500	−0.5	Silicon (very sensitive to purity)	(3.4–15) $\times 10^6$	
Chromium plating	14–66		Silicon bronze (96Cu, 3Si, 1Zn)	21.0	
Constantan (60Cu, 40Ni)	52.5	0.01	Silicon nitride	10^{19}	
Copper	1.678	3.9	Silver	1.6	6.1
Evanohm (75Ni, 20Cr, 2.5Al, 2.5Cu)	134		Sodium	4.75	
Germanium (polycrystalline)	46×10^6		Stainless steel (cast)	70–122	
Gold	2.12	3.4	Tantalum	12.45	3.8
Iridium	5.3		Tantalum carbide	20	
Iron (99.99%)	9.71	6.5	Tin	11.0	4.7
Lead	22	3.36	Titanium	42	
Manganese	185		Titanium and its alloys	48–199	
Manganin	44	0.01	Titanium carbides	105	
Manganin (84Cu, 12Mn, 4Ni)	48		Tungsten	5.6	4.5
Mercury	96	0.89	Zinc	5.9	4.2
Mullite[a]	10^{21}		Zircon[a]	$> 10^{20}$	
Nichrome	100	0.4	Zirconium and its alloys	40–74	
Nickel	6.8	6.9			

[a]Volume resistivity.

Table A.8. Properties of Piezoelectric Materials at 20°C

	PVDF	$BaTiO_3$	PZT	Quartz	TGS
Density ($\times 10^3$ kg/m^3)	1.78	5.7	7.5	2.65	1.69
Dielectric constant, ε_r	12	1700	1200	4.5	45
Elastic modulus (10^{10} N/m)	0.3	11	8.3	7.7	3
Piezoelectric constant (pC/N)	$d_{31} = 20$ $d_{32} = 2$ $d_{33} = -30$	78	110	2.3	25
Pyroelectric constant (10^{-4} C/m^2 K)	4	20	27	—	30
Electromechanical coupling constant (%)	11	21	30	10	—
Acoustic impedance (10^6 kg/m^2 s)	2.3	25	25	14.3	—

Table A.9. Physical Properties of Pyroelectric Materials

Material	Curie Temperature (°C)	Thermal Conductivity (W/mK)	Relative Permitivity (ε_r)	Pyroelectric Charge Coeff. (C/m^2 K)	Pyroelectric Voltage Coeff. (V/mK)	Coupling, k_p^2 (%)
Single Crystals						
TGS	49	0.4	30	3.5×10^{-4}	1.3×10^6	7.5
$LiTa0_3$	618	4.2	45	2.0×10^{-4}	0.5×10^6	1.0
Ceramics						
$BaTiO_3$	120	3.0	1000	4.0×10^{-4}	0.05×10^6	0.2
PZT	340	1.2	1600	4.2×10^{-4}	0.03×10^6	0.14
Polymers						
PVDF	205	0.13	12	0.4×10^{-4}	0.40×10^6	0.2
Polycrystalline Layers						
$PbTiO_3$	470	2 (monocrystal)	200	2.3×10^{-4}	0.13×10^6	0.39

Note: The above figures may vary depending on manufacturing technologies.
Source: From Meixner, H., Mader, G., and Kleinschmidt, P. Infrared sensors based on the pyroelectric polymer polyvinylidene fluoride (PVDF). *Siemens Forsch. Entwicl. Ber. Bd.* 15(3), 105–114, 1986.

Table A.10. Characteristics of Thermocouple Types

Junction Materials	Sensitivity (at 25°C) (μV/°C)	Temperature Range (°C)	Applications	Designation
Copper/constantan	40.9	−270 to 600	Oxidation, reducing, inert, vacuum; preferred below 0°C; moisture resistant	T
Iron/constantan	51.7	−270 to 1000	Reducing and inert atmosphere; avoid oxidation and moisture	J
Chromel/alumel	40.6	−270 to 1300	Oxidation and inert atmospheres	K
Chromel/constantan	60.9	−200 to 1000		E
Pt (10%)/Rh–Pt	6.0	0 to 1550	Oxidation and inert atmospheres; avoid reducing atmosphere and metallic vapors	S
Pt (13%)/Rh–Pt	6.0	0 to 1600	Oxidation and inert atmospheres; avoid reducing atmosphere and metallic vapors	R
Slver–Paladium	10.0	200 to 600		
Constantan–tungsten	42.1	0 to 800		
Silicon–aluminum	446	−40 to 150	Used in thermopiles and micromachined sensors	

Table A.11. Thermoelectric Coefficients and Volume Resistivities of Selected Elements

Element	$\alpha(\mu V\ K^{-1})$	ρ ($\mu\Omega$ m)
p-Si	100–1000	10–500
p-Poly-Si	100–500	10–1000
Antimony (Sb)	32	18.5
Iron (Fe)	13.4	0.086
Gold (Au)	0.1	0.023
Copper (Cu)	0	0.0172
Silver (Ag)	−0.2	0.016
Aluminum (Al)	−3.2	0.028
Ptinum (Pt)	−5.9	0.0981
Cobalt (Co)	−20.1	0.0557
Nickel (Ni)	−20.4	0.0614
Bismuth (Bi)	−72.8	1.1
n-Si	−100 to −1000	10–500
n-Poly-Si	−100 to −500	10–1000

Source: Adapted from Schieferdecker, J., et al. Infrared thermopile sensors with high sensitivity and very low temperature coefficient. *Sensors Actuators A* 46–47, 422–427, 1995.

Table A.11a. Thermocouples for Very Low and Very High Temperatures

Materials	Useful range (°C)	Approx. sensitivity (μV/°C)
Iron–constantan	Down to −272	−32
Copper–constantan	Down to −273	−22.9
Cromel–alumel	Down to −272	−23.8
Tantalum–tungsten	Up to 3000	6.1
Tangsten–tangsten(50)molybdenum	Up to 2900	2.8
Tangsten–tangsten(20)rhenium	Up to 2900	12.7

Table A.12. Densities (kg/m^3) of Some Materials at 1 atm Pressure and 0°C

Best Laboratory Vacuum	10^{-17}	Silica	1,938–2,657
Hydrogen	0.0899	Graphite recrystallized	1,938
Helium	0.1785	Borosilicate glass (TEMPAX®)[a]	2,200
Methane	0.7168	Asbestos fibers	2,400–3,300
Carbon monoxide	1.250	Silicon	2,333
Air	1.2928	Polycrystalline glass	2,518–2,600
Oxygen	1.4290	Aluminum	2,700
Carbon dioxide	1.9768	Mullite	2,989–3,293
Plastic foams	10–600	Silicon nitride	3,183
Benzene	680–740	Alumina ceramic	3,322–3,875
Alcohol	789.5	Zinc alloys	5,200–7,170
Turpentine	860	Vanadium	6,117
Mineral oil	900–930	Chromium	7,169
Natural rubber	913	Tin and its alloys	7,252–8,000
Polyethylene, low density	913	Stainless steel	8,138
Ice	920	Bronzes	8,885
Polyethylene, high density	950	Copper	8,941
Carbon and graphite fibers	996–2,000	Cobalt and its alloys	9,217
Water	1,000	Nickel and its alloys	9,125
Nylon 6	1,100	Bismuth	9,799
Hydrochloric acid (20%)	1,100	Silver	10,491
Acrylics	1,163–1,190	Lead and its alloys	11,349
Epoxies	1,135–2,187	Palladium	12,013
Coal tar	1,200	Mercury	13,596
Phenolic	1,126–2,989	Molybdenum	13,729
Glycerin	1,260	Tantalum and its alloys	16,968
PVC	1,350	Gold	19,320
Saran fibers	1,700	Tungsten and its alloys	19,653
Sulfuric acid (20%)	1,700	Platinum	21,452
Polyester	1,800	Iridium	22,504
Beryllium and its alloys	1,855–2,076	Osmium	22,697

[a]TEMPAX® is a registered trademark of Schott Glasswerke, Mainz, Germany.

Table A.13. Mechanical Properties of Some Solid Materials

Material	Modulus of elasticity (GPa)	Poisson's ratio (ν)	Density (kg/m^3)
Aluminum	71	0.334	2,700
Beryllium copper	112	0.285	8,220
Brass	106	0.312	8,530
Copper	119	0.326	8,900
Glass	46.2	0.125	2,590
Lead	36.5	0.425	11,380
Molybdenum	331	0.307	10,200
Phosphor bronze	11	0.349	8,180
Steel (carbon)	207	0.292	7,800
Steel (stainless)	190	0.305	7,750

Table A.14. Mechanical Properties of Some Crystalline Materials

Material	Yield Strength ($\times 10^{10}$ dyn/cm^2)	Knoop Hardness (kg/mm^2)	Young's Modulus ($\times 10^{12}$ dyn/cm^2)	Density (g/cm^3)	Thermal Conductivity (W/cm °C)	Thermal Expansion ($\times 10^{-6}$/°C)
Diamond[a]	53	7000	10.35	3.5	20.0	1.0
SiC[a]	21	1280	7.0	3.2	3.5	3.3
TiC[a]	20	1270	4.97	4.9	3.3	6.4
Al_2O_3[a]	15.4	2100	5.3	4.0	0.5	5.4
Si_3N_4[a]	14	3486	3.85	3.1	0.19	0.8
Iron[a]	12.6	400	1.96	7.8	0.803	12.0
SiO_2 (fibers)	8.4	820	0.73	2.5	0.014	0.55
Si[a]	7.0	850	1.9	2.3	1.57	2.33
Steel (max. strength)	4.2	1500	2.1	7.9	0.97	12.0
W	4.0	485	4.1	19.3	1.78	4.5
Stainless steel	2.1	660	2.0	7.9	0.329	17.3
Mo	2.1	275	3.43	10.3	1.38	5.0
Al	0.17	130	0.70	2.7	2.36	25.0

[a]Single crystal.

Source: From Petersen, K. E. Silicon as a mechanical material. *Proc. IEEE* 70(5), 420–457, 1982.

Table A.15. Speed of Sound Waves

Medium	Speed (m/s)	Medium	Speed (m/s)
Air (dry at 20°C)	331	Copper	3,810
Steam (134°C)	494	Aluminum	6,320
Hydrogen (20°C)	1,330	Pyrex® glass	5,170
Water (fresh)	1,486	Steel	5,200
Water (sea)	1,519	Beryllium	12,900
Lead	1,190		

Note: Gases at 1 atm pressure, solids in long thin rods

Table A.16. Coefficient of Linear Thermal Expansion of Some Materials (per °C$\times 10^{-6}$)

Material	α	Material	α
Alnico I (permanent magnet)	12.6	Nylon	90
Alumina (polycrystalline)	8.0	Phosphor–bronze	9.3
Aluminum	25.0	Platinum	9.0
Brass	20.0	Plexiglas (Lucite)	72
Cadmium	30.0	Polycarbonate (ABS)	70
Chromium	6.0	Polyethylene (high density)	216
Comol (permanent magnet)	9.3	Silicon	2.6
Copper	16.6	Silver	19.0
Fused quartz	0.27	Solder 50-50	23.6
Glass (Pyrex®)	3.2	Steel (SAE 1020)	12.0
Glass (regular)	9.0	Steel (stainless: type 304)	17.2
Gold	14.2	Teflon	99
Indium	18.0	Tin	13.0
Invar	0.7	Titanium	6.5
Iron	12.0	Tungsten	4.5
Lead	29.0	Zinc	35.0
Nickel	11.8		

Table A.17. Specific Heat and Thermal Conductivity of Some Materials (at 25°C)

Material	Specific Heat (J/kg °C)	Thermal conductivity (W/m °C)	Density (kg/m^3)
Air (1 atm)	995.8	0.012	1.2
Alumina	795	6	4,000
Aluminum	481	88–160	2,700
Bakelite	1,598	0.23	1,300
Brass	381	26–234	8,500
Chromium	460	91	
Constantan	397	22	8,800
Copper	385	401	8,900
Diamond		99–232	
Fiberglass	795	0.002–0.4	60
Germanium		60	
Glass (Pyrex)	780	0.1	2,200
Glass (regular)		1.9–3.4	
Gold	130	296	19,300
Graphite		112–160	
Iron	452	79	7,800
Lead	130	35	11,400
Manganin	410	21	8,500
Mercury	138	8.4	13,500
Nickel and its alloys	443	6–50	8,900
Nylon	1,700	0.12	1,100
Platinum	134	73	21,400
Polyester	1,172	0.57–0.73	1,300
Polyurethane foam		0.012	40
Silicon	668	83.7	2,333
Silicone oil	1,674	0.1	900
Silver	238	419	10,500
Stainless steel	460	14–36	8,020
Styrofoam	1,300	0.003–0.03	50
Teflon TFE	998	0.4	2,100
Tin	226	64	7,300
Tungsten	139	96.6	19,000
Water	4,184	0.6	1,000
Zinc	389	115–125	7,100

Table A.18. Typical Emissivities of Different Materials (from 0°C to 100°C)

Material	Emissivity	Material	Emissivity
Blackbody (ideal)	1.00	Green leaves	0.88
Cavity Radiator	0.99–1.00	Ice	0.96
Aluminum (anodized)	0.70	Iron or steel (rusted)	0.70
Aluminum (oxidized)	0.11	Nickel (oxidized)	0.40
Aluminum (polished)	0.05	Nickel (unoxidized)	0.04
Aluminum (rough surface)	0.06–0.07	Nichrome (80Ni–20Cr) (oxidized)	0.97
Asbestos	0.96	Nichrome (80Ni–20Cr) (polished)	0.87
Brass (dull tarnished)	0.61	Oil	0.80
Brass (polished)	0.05	Silicon	0.64
Brick	0.90	Silicone rubber	0.94
Bronze (polished)	0.10	Silver (polished)	0.02
Carbon-filled latex paint	0.96	Skin (human)	0.93–0.96
Carbon lamp black	0.96	Snow	0.85
Chromium (polished)	0.10	Soil	0.90
Copper (oxidized)	0.6–0.7	Stainless steel (buffed)	0.20
Copper (polished)	0.02	Steel (flat rough surface)	0.95–0.98
Cotton cloth	0.80	Steel (ground)	0.56
Epoxy Resin	0.95	Tin plate	0.10
Glass	0.95	Water	0.96
Gold	0.02	White paper	0.92
Gold black	0.98–0.99	Wood	0.93
Graphite	0.7–0.8	Zinc (polished)	0.04

Table A.19. Refractive Indices (n) of Some Materials

Material	n	Wavelength (μm)	Note
Vacuum	1		
Air	1.00029		
Acrylic	1.5	0.41	
AMTIR-1	2.6	1	Amorphous glass[a]
($Ge_{33}As_{12}Se_{55}$)	2.5	10	
AMTIR-3	2.6	10	Amorphous glass[a]
($Ge_{28}Sb_{12}Se_{60}$)			
As_2S_3	2.4	8.0	Amorphous glass[a]
CdTe	2.67	10.6	
Crown glass	1.52		
Diamond	2.42	0.54	Excellent thermal conductivity
Fused silica (SiO_2)	1.46	3.5	
Borosilicate glass	1.47	0.7	TEMPAX[b] Transparent: 0.3–2.7μm
GaAs	3.13	10.6	Laser windows
Germanium	4.00	12.0	
Heaviest flint glass	1.89		
Heavy flint glass	1.65		
Irtran 2 (ZnS)	2.25	4.3	Windows in IR sensors
KBr	1.46	25.1	Hygroscopic
KCl	1.36	23.0	Hygroscopic
KRS-5	2.21	40.0	Toxic
KRS-6	2.1	12	Toxic
NaCl	1.89	0.185	Hygroscopic, corrosive
Polyethylene	1.54	8.0	Low-cost IR windows/lenses
Polystyrene	1.55		
Pyrex 7740	1.47	0.589	Good thermal and optical properties
Quartz	1.54		
Sapphire (Al_2O_3)	1.59	5.58	Chemically resistant
Silicon	3.42	5.0	Windows in IR sensors
Silver bromide (AgBr)	2.0	10.6	Corrosive
Silver chloride (AgCl)	1.9	20.5	Corrosive
Water (20°C)	1.33		
ZnSe	2.4	10.6	IR windows, brittle

[a]Available from Amorphous Materials, Inc., Garland, TX.
[b]TEMPAX® is a registered trademark of Schott Glasswerke, Mainz, Germany.

Table A.20. Characteristics of C–Zn and Alkaline Cells

Battery	W h/L	W h/kg	Drain Rate	Shelf Life
Carbon–zinc	150	85	Low–medium	2 years
Alkaline	250	105	Medium–high	5 years

Source: From Powers, R.A. Batteries for low power electronics. *Proc. IEEE*, 83(4), 687–693, 1995.

Table A.21. Lithium–Manganese Dioxide Primary Cells

Construction	Voltage	Capacity (mA h)	Rated dc Current (mA)	Pulse Current (mA)	Energy Density (W h/L)
Coin	3	30–1,000	0.5–7	5–20	500
Cyl. wound	3	160–1,300	20–1,200	80–5,000	500
Cyl. bobbin	3	650–500	4–10	60–200	620
Cyl. "D" cell	3	10,000	2,500		575
Prismatic	3	1,150	18		490
Flat	3/6	150–1,400	20–125		290

Source: From Powers, R.A. Batteries for low power electronics. *Proc. IEEE*, 83(4), 687–693, 1995.

Table A.22. Typical Characteristics of "AA"-Size Secondary Cells

System	Volts	Capacity (mA h)	Rate[a] (*C*)	W h/L	W h/kg	Cycles	Loss/month (%)
NiCad	1.2	1000	10	150	60	1000	15
Ni–MH	1.2	1200	2	175	65	500	20
Pb acid	2	400	1	80	40	200	2
Li ion (CoO_2)	3.6	500	1	225	90	1200	8
Li/MnO_2	3	800	0.5	280	130	200	1

[a] *Note*: Discharge rate unit, *C* (in mA), is equal numerically to the nominal capacity (in mA h).

Table A.23. Miniature Secondary Cells and Batteries

Manufacturer	Part	Type	Size	Capacity (mA h)	Voltage	Price $ (approx)
Avex Corp., Bensalem, PA (800-345-1295)		RAM	AA	1.4	1.5	1
GN National Electric Inc., Pomona, CA (909-598-1919)	GN-360	NiCd	15.5 × 19 mm	60	3.6	1.10
GP Batteries USA, San Diego, CA (619-674-5620)	Green-Charge	NiMH	2/3AA, AA, 2/3AF, 4/5AF	600–2500	1.2	2–7
Gould, Eastlake, OH (216-953-5084)	3C120M	$LiMnO^2$	3 × 4 × 0.12 cm	120	3	2.71
House of Batteries Inc., Huntington Beach, CA (800-432-3385)	Green cell	NiMH	AA, 4/5A, 7/5A	1200–2500	1–2	3.50–12
Maxell Corp., Fairlawn, NJ (201-794-5938)	MHR-AAA	NiMH	AAA	410	1.2	4
Moli Energy Ltd., Maple Ridge, BC, Canada (604-465-7911)	MOLICEL	Li ion	18(diam)×65 mm	1200	3.0–4.1	25
Plainview Batteries, Inc., Plainview, NY (516-249-2873)	PH600	NiMH	48 × 17 × 7.7 mm	600	1.2	4
Power Coversion, Inc., Elmwood Park, NJ (201-796-4800)	MO4/11	$LiMnO^2$	1/2AA	1000	3.3	5–8
Power Sonic Corp., Redwood City, CA (415-364-5001)	PS-850AA	NiCd	AA	850	1.2	1.75
Rayovac Corp., Madison, WI (608-275-4690)	Renewal	RAM	AA, AAA	1200, 600	1.5	From 0.50
Renata U.S., Richardson, TX (214-234-8091)	CR1025	Li	10 mm	25	3.0	0.50

Table A.23 *Continued*

Manufacturer	Part	Type	Size	Capacity (mA h)	Voltage	Price $ (approx)
Sanyo Energy (U.S.A.), San Diego, CA (691-661-7992)	Twicell	NiMH	10.4 × 44.5 × 67 mm	450	1.2	3.85
Saft America, Inc., San Diego, CA (619-661-7992)	VHAA	NiMH	AA	1100	1.2	2.95
Tadiran Electronics, Port Washington, NY (516-621-4980)		Li	1/AA–DD packs	370 mAh to 30 Ah	3–36	1+
Toshiba America, Deerfield, IL (800-879-4963)	LSQ8	Li ion	8.6 × 3.4 × 48 mm	900	3.7	12–15
Ultralife Batteries, Inc., Newark, NJ (315-332-7100)	U3VL	Li	25.8 × 44.8 × 16.8	3600	3.0	4.60
Varta Batteries, Inc., Elmsford, NY (914-592-2500)		NiMH	AAA-F	300–8,000	1.2	0.80+

Note: Li ion = lithium ion, $LiMnO^2$ = lithium manganese dioxide, NiCd = nickel–cadmium, NiMH = nickel–metal hydride, RAM = rechargable alkaline manganese.

Table A.24. Electronic Ceramics (Between 25°C and 100°C)

Alumina	96% (BeO) (Al_2O_3)	Beryllia Nitrade	Boron Nitrade (BN)	Aluminum Carbide (AlN)	Silicon (SiC)	Silicon (Si)
Hardness, Knopp (kg/mm^2)	2000	1000	280	1200	2800	—
Flexural strength (10^5 N/m^2)	3.0	1.7–2.4	0.8	4.9	4.4	—
Thermal conductivity (W/(m K)	21	250	60	170–200	70	150
Thermal expansion (10^{-6}/K)	7.1	8.8	0.0	4.1	3.8	3.8
Dielectric strength (kV/mm)	8.3	19.7	37.4	14.0	15.4	—
Dielectric loss (10^{-4} tan delta at 1 MHz)	3–5	4–7	4	5–10	500	—
Dielectric constant, κ (at 10 MHz)	10	7.0	4.0	8.8	40	—

Table A.25. Properties of Glasses

	Soda-Lime	Borosilicate	Lead glass	Alumosilicate	Fused Silica	96% Silica
Modulus of elsticity (10^6 psi)	10.2	9.0	8.5–9.0	12.5–12.7	10.5	9.8
Softening temperature (°F)	1285	1510	932–1160	1666–1679	2876	2786
Coefficient of thermal expansion (10^{-6} in./in. °C)	8.5–9.4	3.2–3.4	9–12.6	4.1–4.7	0.56	0.76
Thermal conductivity (BTU—in./h ft^2 °F)	7.0	7.8	5.2	9.0	9.3	10.0
Density (lbs/in^3)	0.089	0.081	0.103–0.126	0.091–0.095	0.079	0.079
Electrical resistivity (log10Ω cm)	12.4	14	17	17	17	17
Refractive index	1.525	1.473	1.540–1.560	1.530–1.547	1.459	1.458

This page intentionally left blank

Index

α-particles, 443

A/D, 175, 177, 183
aberrations, 136
ablation sensor, 293, 294
ABS, 539
absolute sensor, 7, 461
absolute temperature, 507
absolute zero, 96
absorber, 107
absorption, 127, 145, 514
absorption coefficient, 129
absorptivity, 133
acceleration, 217, 218, 301, 312
accelerometer, 29, 113, 116, 304
accuracy, 17, 219, 305, 398
acoustic, 73, 146, 331
acoustic measurements, 383
acoustic noise, 228
acoustic pressure, 93
acoustic sensors, 381, 388
acoustics, 69, 381
acousto-optic, 147
acrylic, 539
active bridge, 200
active sensor, 7, 164
actuator, 75
additive noise, 208
AFIR, 426, 477
aging, 29
air bubble, 257
airflow, 375
aluminum, 400, 430, 550
aluminum coatings, 541
aluminum nitride, 75
aluminum oxide, 400
AM, 290
Ampere's law, 54
amperometric devices, 503
amplifier, 155, 186, 416
AMTIR, 543
angular displacement, 316
angular encoding, 270
antenna, 173, 290
aperture, 142
appliances, 227
arsenic trisulfate, 543
ASIC, 156
attenuation coefficient, 129
auxiliary electrode, 505
avalanche, 449, 450, 455
avalanche photodiodes, 414

band gap, 408
band-gap references, 171
bandwidth, 27, 205
barometer, 342
battery, 222
bead-type thermistors, 473
beams, 318
Becquerel, 89
becquerel, 444
Beer's law, 108
Beer–Lambert law, 515
Bell, 93
bellows, 342
Bernoulli, 339, 361
beryllium, 541

bias current, 154, 156
bias resistor, 251
bimetal, 97
binary codes, 176
biological sensors, 388, 519
bismuth, 451
blackbody, 106, 109, 130
bolometer, 434, 435
Boltzmann constant, 205, 413
bonding, 554
boron, 552
Boyle, 339
brass, 144
breakwire, 294
breeze sensor, 374
bridge circuit, 401
brightness, 130
British unit of heat, 95
broadband detectors, 426

cable, 213
cadmium telluride, 454
calibration, 18, 25, 466
calibration error, 19
calibration temperature, 98, 470
Callendar–van Dusen, 463
candela, 130
cantilever, 372
cantilever beam, 330
capacitance, 44, 69, 187, 211, 233, 259, 387, 399, 415, 508
capacitive accelerometer, 306
capacitive bridge, 261
capacitive coupling, 234
capacitive sensor, 48, 161, 162, 350, 396
capacitor, 44, 67, 76, 78, 155, 162, 167, 172, 178, 187, 212, 234, 238, 261, 306
catalytic devices, 510
cavity, 19, 278, 311, 373, 426
cavity effect, 109
CdS, 421
Celsius, 96, 491
ceramic, 309, 542
characteristic temperature, 64, 467
charge, 188
charge amplifier, 161
charge detector, 234
charge-balance, 179
charge-to-voltage converter, 188
chemFET, 504
chemical poisoning, 501
chemical reaction, 512, 513
chemical sensor, 3, 142, 499
chemical species, 505
chemiluminescence, 514
chemometrics, 521, 523
chip thermistor, 473
chromatogram, 521
circuit protection, 480
cladding, 140, 142, 516
Clark electrode, 508
clock, 183
CMOS, 310, 374, 436
CMRR, 156, 160, 210
CO, 506
CO_2, 108, 506, 514, 515
coating, 135
coaxial cables, 174
cobalt, 55, 63
coefficient of reflection, 126
coil, 57, 255, 266, 302, 319
cold junction, 89
collector, 418
comparator, 171, 181
complex devices, 502
complex sensor, 4, 512
concentrator, 144
condenser microphones, 382
conductance, 504
conduction, 99
conduction band, 408
conductive plastics, 540
conductivity, 61, 409, 503
conductivity sensor, 507
conductometric devices, 503
constantan, 481
contact resistance, 101
contact sensor, 457
contamination, 500
convection, 99, 102
converter, 177, 431
copolymer, 75
copper, 38, 60, 111, 135, 385, 540, 541
Coriolis, 376
Coriolis acceleration, 314
Coriolis force, 316
Coriolis tube, 376
cost, 161

Coulomb's law, 40
cross-talk, 215
crystal, 76, 147, 335, 389, 409, 421, 538, 550
crystalline materials, 42, 408
Curie point, 77
Curie temperature, 32, 70, 80, 266, 320, 477
current, 511
current generator, 162, 165
current mirror, 165
current pump, 165
current sink, 165
current source, 165, 199, 203
cutoff frequency, 26
CVD, 76, 372, 546

D/A, 176
DAC, 189
damping, 28
damping factor, 28
damping medium, 304
Darlington connection, 419
data acquisition, 151
dead band, 23
decibel, 93
deflection, 13
dew point, 393, 402
Dewar, 296
Dewar cooling, 423
diaphragm, 253, 331, 342, 349, 382, 386
dielectric, 46, 386
dielectric absorption, 215
dielectric constant, 46, 297, 398
differential equation, 26, 116
differential sensor, 210
diffusion, 552
digital format, 177
diode, 488
diode sensor, 489
dipole, 46
dipole moment, 42, 78
direct conversion, 38
direct devices, 502
direct sensor, 3, 4
disbalanced bridge, 193
displacement, 253, 271, 274, 301
displacement sensor, 285
dissipation constant, 478
dissipation factor, 475
distance sensor, 270
distortion mask, 241
divider, 190
door openers, 231
Doppler, 230
Doppler effect, 229, 287, 367
drag element, 377
drag force sensor, 377
driven shield, 235
dual-ramp, 175
dual-slope, 181
dynamic error, 25
dynamic range, 15, 320, 367
dynodes, 445

eddy currents, 264, 292
Einstein, 407
elasticity, 92, 387
electret, 386
electret microphone, 387
electric charge, 38, 39, 59, 331
electric current, 83, 451
electric dipole, 42
electric field, 39, 40, 59, 418, 446
electric potential, 43
electrical conduction, 60
electrochemical cell, 506, 508
electrochemical sensor, 505, 526
electrode, 236, 237, 260, 296, 508
electrolyte, 500, 506, 509
electromagnetic flowmeter, 370
electromagnetic radiation, 238, 407
electromagnetic sensor, 302
electrometer, 507
electromotive force, 56
electron, 51, 82, 408
electron multiplication, 447
electron–hole pairs, 451
electronic nose, 499, 526
electrostatic, 212
electrostatic gyro, 314
electrostatic shield, 213
emissivity, 105, 106, 146, 245, 430
emitter, 107, 248, 418
encoding disk, 282
energy bonds, 537
e-nose, 524, 525
enzyme, 513
enzyme sensors, 520
epoxy, 29, 308, 540

error, 33, 186
etch mask, 550, 553
etching, 549
Euler, 323
excimer laser, 547
excitation, 25, 493
excitation signal, 7
excitation voltage, 204

Fabry–Perot, 149, 278, 353, 427
Fahrenheit, 95
failure, 31
farad, 45
Faraday, 52, 56, 370
Faraday cage, 41
Faraday's Law, 52, 302
Faradic current, 511
far-infrared, 107, 111, 132, 135, 425
far-infrared (AFIR), 437
feedback, 4, 161, 167, 205, 335
Ferdinand II, 95
Fermi, 89
ferroelectric, 66, 477
ferromagnetic, 52, 263
FET, 500
fiber, 140, 383
fiber-optic, 147, 275, 436, 515
fiber-optic sensor, 142, 278
field lines, 39, 45
filament, 479
film, 399, 435
film transducers, 75
filter, 181, 279
filtering, 124
first-order response, 114
flame, 439
flow, 369
flow measurement, 361
flow rate, 360
flow resistance, 361
flowmeter, 348, 366
fluid, 329, 339, 383
fluoroplastics, 539
fluoroptic method, 493
flux, 40, 109, 126, 130, 144, 250
focus, 136
focusing lens, 420
foil, 487
follower, 155
force, 39, 323, 327, 333, 334, 377
forced convection, 102
format, 37
Fourier, 91
Fourier transforms, 522
FP interferometer, 353
FPA, 435
Fraden Model, 468
Franklin, 38
frequency, 517
frequency range, 27
frequency response, 26, 153
Fresnel, 138
Fresnel lens, 138, 247
frost point, 402
FSR, 332
FTIR, 522
full scale, 15

gain–bandwidth product, 157
Galileo, 339
γ-radiation, 443
γ-rays, 103
gas, 4, 108, 309, 333, 341, 354, 363, 374, 376, 439, 449, 499, 514, 537
gas analyzer, 425
gas chromatographs, 520
gas sensor, 503, 510, 524
gauge, 65
gauge sensor, 348
Gauss' law, 40, 41
Gaussian System, 9
Geiger–Müller counter, 450
geometrical optics, 123
geometry factor, 47
germanium, 39, 436, 454
Gilbert, 50
glass, 140, 509, 543, 554
glucose, 509
Golay cells, 426
gold, 135, 144, 542
gold black, 146
GPS, 301
grating, 281
gravimetric detector, 517
gravitational sensor, 256
gravity, 305
Gunn oscillator, 229
gyroscope, 313

H_2O, 108
Hall, 82
Hall coefficient, 83
Hall effect, 82, 267, 346, 534
Hall effect sensor, 85, 268
harmonic, 230, 334
heat, 457
heat absorption, 475
heat capacity, 98
heat loss, 364
heat pump, 402
heat sink, 81, 309
heat transfer, 99
heated probe, 520
heat-flow detector, 77
Henry, 56, 370
Hooke, 95
hot junction, 89
Howland, 167
humidity, 29, 35, 75, 393, 526
humidity sensor, 49, 401
hybrid, 159
hydrocarbon fuel, 512
hydrocarbon sensor, 500
hydrogel, 505
hydrophone, 381
hygristors, 66
hygrometer, 402
hygrometric sensor, 399
hysteresis, 20, 253, 332

identification, 499
illumination, 130
image, 242
immobilization, 519
inclination detectors, 256
index of refraction, 125, 141
inductance, 57
inertia, 26, 115
inertial mass, 308, 311
infrared, 105, 238, 434, 492, 514
infrared detectors, 425
infrared flux, 3
infrared sensor, 543
infrasonic, 388
inherent noise, 204
input, 151
input impedance, 151, 152, 158
input resistance, 154
input stage, 152
instrumentation amplifier, 159
insulation, 486
integrator, 179
intensity sensor, 143
interface circuit, 152
interferometer, 280, 383
intrusion, 173
intrusive sensors, 294
ion, 448
ionization, 447
ionizing chamber, 448
ionizing radiation, 29, 164, 450
IR, 128
IR detector, 107
IR spectrometers, 520
iridium, 542
ISA, 481
ITS-90, 463

JFET, 154, 219, 375, 388, 431, 432
Johnson noise, 205, 206
Joule, 89, 355, 364
Joule heat, 479
junction, 427, 482
junction capacitance, 411

Kawai, 72
Kelvin, 7, 96, 248, 490, 496
keyboard, 324
kinetic energy, 60, 104
Kirchhoff, 61, 105, 145, 458
Kirchhoff's laws, 61, 118
KOH etch, 550
Korotkoff sounds, 385
krypton, 451
KTY, 464

Laplace transforms, 303
laser, 112, 384
laser gyro, 318
law of reflection, 125
LC, 171
LCD, 113
lead, 541
leakage current, 154, 414
least squares, 22
LeChatelier, 89
LED, 37, 195, 239, 258, 280, 283, 494, 515

lens, 240, 242, 249, 277, 420
Leslie, 393
level detectors, 278, 291
life test, 32
LIGA, 547
light, 111, 123, 136, 146, 243, 276, 411, 445, 494, 511
linearity, 21
liquid, 278, 296, 363
lithium, 66, 147, 222, 453
load cells, 324
logarithmic scale, 15
logic circuits, 171
long-term stability, 29
loudspeaker, 3
lumen, 130
luminescence, 441
LVDT, 263, 302, 325

magnesium, 541
magnet, 55, 268, 274
magnetic field, 50, 53, 82, 103, 215, 268, 314, 317, 371, 447
magnetic flux, 267, 351
magnetic noise, 216
magnetic pole, 50
magnetic reluctance sensor, 275
magnetic sensor, 262, 540
magnetic shielding, 216, 540
magnetism, 50
magnetite, 50
magnetization, 52
magnetoresistive sensor, 271
magnetostrictive detector, 274
manganese, 63
mass, 305, 324, 359, 378, 516
mass spectrometers, 520
material characteristic, 468
matrix, 487
Maxwell, 371
MCT, 423
measurand, 2
membrane, 427, 525
MEMFET, 505
MEMS, 269, 427, 430, 439, 534, 547
MEMSIC, 310
MEOMS, 547
mercury switch, 257
metal, 223, 272, 408, 504, 540
metal carbides, 542
metal films, 145, 549
metal oxide, 69, 473, 503
metallic electrode, 452
metallization, 473
Michelson, 383
microbalance method, 525
microbalance sensors, 516
microcalorimetry, 513
microcontroller, 185
microgravimetric technique, 517
micromachining, 547, 549
microphone, 381
microsensor, 510
microwave, 288, 477
microwave devices, 434
mid-infrared, 111, 425, 426
military standard, 32
MIR, 289
mirror, 134, 242
modulation, 171, 290
moisture, 66, 393, 396
molybdenum, 542
monolithic sensors, 491
MOS, 188
motion detector, 136, 227, 237
MTBF, 31
multiplexing, 183
multiplicative noise, 209
multivibrator, 178
mutual inductance, 58

natural frequency, 27, 344
near-infrared, 111, 420
Nernst equation, 507, 511
neural network, 527
neuron, 528
Newton, 95, 115, 126
Newton's law, 313
nichrome, 326, 430, 487
nickel, 63, 224, 541
noise, 178, 204, 219, 238, 275, 312, 351, 417, 423, 522
nonlinearity, 20, 21
NTC, 465, 503
nuclear radiation, 443
n-wells, 85
nylon, 539
Nyquist, 371

occupancy sensors, 227
odor classification, 528
odor sensor, 524
Oersted, 51
offset, 186
offset voltage, 153, 156
Ohm's law, 100, 162, 165, 432
olfactory cells, 525
one-shot, 179
OPAM, 156, 172, 188
open-loop, 157
open-loop gain, 188
operational amplifier, 156, 201
optical cavity, 353
optical contrast, 240
optical detection, 412
optical modulation, 514
optical paths, 516
optical power, 412
optical sensor, 494, 514, 520
optocoupler, 403
organic, 29
oscillating hygrometer, 403
oscillating response, 28
oscillating sensor, 516
oscillator, 172, 178, 187, 235, 263, 390, 396
output capacitance, 154
output current, 167
output impedance, 24
output resistance, 165
output signal format, 2
oxygen, 67, 499, 508, 509

p-n junction, 18, 284, 353, 408, 411, 488
palladium, 542
parallel-plate capacitor, 45
parametric methods, 523
Pascal, 339
passive sensor, 7
Pellister, 514
Peltier, 90, 403
Peltier effect, 90, 423, 438
pH, 515
phase, 1, 305
phase lag, 153
phase shift, 27
phenolic, 540
phosphor, 492
photocatalytic sensors, 511
photocathode, 446
photocurrent, 421
photodetector, 131, 276, 410, 419, 445
photodiodes, 410, 411
photoeffect, 37, 284, 407
photoelectron, 445
photomultiplier, 407, 422, 444
photon, 13, 112, 407, 445, 450
photoresist, 148, 548
photoresistor, 195, 243, 410, 420
photosensor, 419
phototransistors, 410, 418
photovoltaic mode, 414, 415
piecewise approximation, 14
piezoelectric, 245, 288, 309, 319, 320, 324, 334, 368, 385, 389, 517
piezoelectric crystal, 431, 496
piezoelectric effect, 66, 496, 518, 534
piezoelectric film, 320, 328, 543
piezoelectric hygrometer, 403
piezoelectric plastics, 540
piezoresistive accelerometer, 307
piezoresistive bridge, 308
piezoresistive effect, 64, 325
piezoresistive gauge, 344
piezoresistive sensors, 350
PIN photodiode, 413
pink noise, 206
pipe, 61
PIR, 145, 245, 426, 427, 430, 437
Pirani gauge, 354
Planck, 103
Planck's constant, 407
Planck's law, 103
plano-convex lens, 136
plastic, 112, 140, 231, 331, 536
platinum, 63, 64, 145, 461, 487, 505, 514, 542
platinum film, 436
Poisson ratio, 92
polarization, 46, 71, 79, 112, 147, 276
polarization filter, 112, 277
poling, 43, 70
polycarbonate, 539
polyester, 539
polyethylene, 107, 155, 536, 539
polymer, 74, 140, 332, 396
polymer films, 80, 512
polymer matrix, 512, 525

polymerization, 538
polypropylene, 539
polysilicon, 430, 436, 535, 550
polystyrene, 399
polyurethane, 539
popcorn noise, 205
position, 253, 270
position-sensitive detector, 283
potential, 511
potentiometer, 255
potentiometric devices, 503, 507
preaging, 474
predictive, 457
pressure, 189, 324, 339
pressure gradient, 361
pressure sensor, 6, 197, 227, 280, 341, 350, 373, 381
primary cells, 223
prototype, 219
proximity, 234
proximity detector, 277
proximity sensor, 253, 260
PS, 536
PSD, 281, 283, 427
p-substrate, 85
PTC thermistor, 477
PVDF, 71, 72, 247, 320, 328, 385
PWM, 189
pyroelectric, 76, 245, 430
pyroelectric coefficient, 247
pyroelectric current, 433
pyroelectric sensor, 30, 73, 76, 161
pyroelectricity, 77
pyrometry, 425
PZT, 389, 434

Q-spoilers, 175
quantification, 499
quantum detector, 161, 407, 423
quartz, 67, 403, 516, 534
quenching, 450

radar, 6, 289, 297
radiation, 95, 99, 111, 133, 239, 448, 451, 457
radiation bandwidth, 104
radiation detection, 447
radiation spectrum, 103
radio waves, 291
radio-frequency, 172
radioactivity, 443
ratiometric technique, 190
RC, 171
reactive ion etching, 553
redox reactions, 506
reference, 190, 200
reference diode, 170
reference electrode, 505, 510
reference sensor, 484
reference temperature, 62, 485
reference voltage, 187
reflection, 124
reflective surface, 134
reflectivity, 133
refraction, 124
refractive index, 141, 147
relative humidity, 49, 394, 396
relative sensors, 461
reliability, 31
Renaldi, 95
repeatability, 23
resistance, 59, 254, 284, 341, 535
resistance multiplication, 164
resistive bridge, 193
resistive load, 165
resistive sensor, 153
resistivity, 60
resistor, 61, 164, 179, 190, 203, 247, 344, 349, 372, 375, 465, 476
resolution, 23, 181, 186, 205, 316, 374
resonant, 304, 387
resonant sensors, 553
resonator, 317
retina, 144
return electrode, 505
Reyleigh waves, 517
RF, 229
RH, 32
rhodium, 542
roentgen, 444
root-sum-of-squares, 34
rotor, 313
RTD, 355, 366, 477, 481, 514
RVDT, 264

Sagnac effect, 317
sampling rate, 177
saturation, 22

SAW, 75, 388, 404, 496, 501, 517
scale, 95
Schmitt trigger, 267
Schottky noise, 206
scintillation counters, 444
secondary cells, 224
second-order response, 114
security alarms, 231
security system, 234
Seebeck, 86, 484
Seebeck coefficient, 87
Seebeck effect, 3, 221, 534
Seebeck potential, 88, 483
selectivity, 500, 524
self-heating, 467, 474
self-heating error, 30
self-heating sensor, 366
self-induction, 57
semiconductor, 408, 421, 484
semiconductor detectors, 451
semiconductor diode, 452
sensitivity, 14, 25, 194, 304, 312, 387, 422, 500, 524
SFB, 347
shield, 235, 260
shielding, 212
signal conditioning, 151
signal-to-noise ratio, 156
signatures, 527
silicate, 543
silicon, 39, 85, 197, 284, 306, 317, 344, 349, 353, 398, 464, 489, 504, 517, 533
silicon bonding, 549
silicon diaphragms, 550
silicon diode, 170, 452
silicon dioxide, 549
silicon micromachining, 547
silicon nitride, 549
silicon plate, 517
silicon sensor, 464
silicon wafer, 548, 550, 554
silicone, 102, 494
silicone oil, 304
silkscreen, 327
silver, 70, 542
SiO_2, 413
skin, 112
smart chemical sensors, 530
Snell, 125
Snell's law, 125, 141
SnO_2, 503
solder, 221
solenoid, 54
solid-state detectors, 455
sound waves, 92
span errors, 199
species, 499
specific heat, 82, 98
specific resistivity, 61, 409
spectroscopy, 449
spectrum, 112, 213
speed, 301
speed response, 26
spherical mirrors, 136
spin-casting, 544
spinning-rotor gauge, 356
sputtering, 545, 546
square-wave oscillator, 171
statistical methods, 523
Stefan–Boltzmann constant, 105, 249, 372
Stefan–Boltzmann law, 14, 105, 244, 249, 372, 429, 437
Steinhart and Hart model, 470
stimulus, 2, 153, 190, 202, 209
storage, 29
straight line, 18
strain, 13, 65, 326, 342, 377
strain gauge, 143, 324, 325, 332
stress, 343
stress detectors, 227
string, 92
substrate, 517
successive-approximation technique, 175
supervised classification, 523
switched-capacitor, 187
synchronous detector, 263
systematic error, 524
systematic inaccuracy, 17

tactile sensor, 327
target species, 500
TCR, 199, 479
Teflon, 155, 388, 509
temperature, 30, 47, 60, 75, 77, 88, 94, 129, 169, 195, 206, 219, 244, 310, 335, 355, 363, 398, 425, 435, 457, 476, 525
temperature coefficient, 159, 280
temperature compensation, 196

temperature correction, 49
temperature differential, 367
temperature gradient, 117
temperature profile, 101
temperature sensitivity, 536
temperature sensor, 19, 49, 312, 403, 460, 490
TGS, 80
thermal accelerometer, 309
thermal capacitance, 26, 81, 98
thermal conductivity, 100, 355, 401, 460
thermal coupling, 436
thermal expansion, 96
thermal feedback, 481
thermal flux, 106, 438
thermal grease, 481
thermal mass, 427
thermal radiation, 14, 78, 144, 244, 426, 434
thermal resistance, 81, 117, 458
thermal shock, 33
thermal time constant, 435
thermistor, 5, 35, 62, 401, 435, 465, 513
thermoanemometer, 363
thermochromic solution, 494
thermocouple, 7, 89, 311, 427, 481
thermocouple amplifier, 485
thermocouple assemblies, 486
thermocouple loop, 482
thermodynamics, 513
thermoelectric, 438
thermoelectric coefficients, 429
thermoelectric coolers, 423
thermoelectric law, 482
thermoelectric voltage, 484
thermoelectricity, 87
thermometer, 109, 403, 468
thermopile, 89, 311, 427, 484
thermoplastic, 538
thermostat, 257, 480
thermowell, 486
thick films, 465
thick oxide, 552
thickness, 293
thin film, 487, 517, 554
thin plate, 344
thin-film material, 544
Thompson, 86
Thomson heating, 91
threshold, 186
threshold circuits, 240
threshold device, 171
tilt sensor, 257
time constant, 26, 81, 118, 460, 476, 492
TiO_2, 511
titanium, 63
toroid, 55
torque, 43, 313
Torricelli, 339
total internal reflection, 141
transceiver, 229
transducer, 3
transfer function, 13, 17, 19, 29, 210, 525
transistor, 418, 488
transition temperature, 478
transmission, 148
transmittance, 128, 133
transmitted noise, 208, 211, 228
triboelectric detectors, 237
triboelectric effect, 38
true value, 13
tube, 144
tube of flow, 359
tungsten, 62, 542
two-wire transmitter, 202
two-point calibration, 19

U.S. Customary System, 9
ultrasonic, 274, 367, 385
ultrasonic crystal, 496
ultrasonic waves, 287, 496
ultraviolet (UV), 111, 439, 492, 511, 543, 548
uncertainty, 18, 33
unsupervised classification, 523

V/F, 176, 177
vacuum, 46, 111, 135, 146, 314, 333, 354, 400
vacuum chamber, 544, 545
vacuum deposition, 545
vacuum sensor, 356
vacuum tube, 356
valence band, 409, 421
VCR, 433
vector, 53, 82
vehicle, 301
velocity, 301, 359, 362, 368
velocity of light, 111

velocity sensor, 302
vertex curvature, 140
vibrating gyro, 314
vibration, 303
vibration detectors, 227, 331
virtual ground, 163, 167, 432
Volta, 222
voltage follower, 158, 432
voltage offset, 205
voltage source, 202
voltage-to-current converter, 202
voltage-to-frequency (V/F), 175
Voltaic pile, 51
voltammetry, 522
VRP, 351

Warburg impedance, 508
warm-up time, 25
warping, 97
water, 386
water tank, 48
water-level sensor, 48
waveguide, 143, 144, 148, 232, 275
wavelength, 104, 126, 410
weber, 55
Wheatstone bridge, 192, 195, 204, 273, 275, 341, 344, 513, 514
white noise, 206
Wiedemann effect, 274
Wien's law, 104
window, 132
window comparator, 240
wiper, 254
wire, 51
work function, 408
working electrode, 505

xenon, 451, 492
X-rays, 443, 547

Young's modulus, 73, 92

zener diode, 169
zinc, 542
zinc oxide, 75